Centre for Efficiency and Productivity Analysis

Working Paper Series
No. WP162021

Performance Analysis: Economic Foundations & Trends

Valentin Zelenyuk

Date: October 2021

School of Economics
University of Queensland
St. Lucia, Qld. 4072
Australia

ISSN No. 1932 - 4398

Performance Analysis:

Economic Foundations & Trends

Valentin Zelenyuk[1]

2021

[1]School of Economics and Centre for Efficiency and Productivity Analysis, The University of Queensland, Australia. e-mail: v.zelenyuk@uq.edu.au.
This is a working paper version that was accepted and (after further editing refinements) was published and to be cited as: Valentin Zelenyuk (2021), "Performance Analysis: Economic Foundations and Trends", Foundations and Trends® in Econometrics: Vol. 11: No. 3, pp 153-229. http://dx.doi.org/10.1561/0800000034 (posted here as CEPA WP with the publisher's agreement).

Contents

Abstract

The goal of this monograph is to very concisely outline the economic theory foundations and trends of the field of Efficiency and Productivity Analysis, also sometimes referred to as Performance Analysis. We start with the profit maximization paradigm of mainstream economics, use it to derive a general profit efficiency measure and then present its special cases: revenue maximization and revenue efficiency, cost minimization and cost efficiency. We then consider various types of technical and allocative efficiencies (directional and Shephard's distance functions and related Debreu-Farrell measures as well as non-directional measures of technical efficiency), showing how they fit or decompose the profit maximization paradigm. We then cast the efficiency and productivity concepts in a dynamic perspective that is frequently used to analyze the productivity changes of economic systems (firms, hospitals, banks, countries, etc.) over time. We conclude this monograph with an overview of major results on aggregation in productivity and efficiency analysis, where the aggregate productivity and efficiency measures are theoretically connected to their individual analogues.

Introduction

The goal of this monograph is to give a relatively concise overview of Efficiency and Productivity Analysis—a very important field of research and practice, spanning over and engaging with many disciplines, most prominently Economics (theoretical and applied), Statistics (and therefore Econometrics), Operations Research (OR) and Management Science (MS), as well as Business Analytics and Business Information Systems, Computer Science and Engineering, etc. Methods developed in this field became very popular in practice for analyzing the efficiency of various economic systems: firms or its distinct departments, branches or plants, entire industries or sub-industries, countries, regions or provinces, as well as various groups or unions of countries, such as APEC, EU, OECD, etc.

While many very good articles and books have been written for this fascinating field, often they were either focusing on certain aspects (in case of articles), which made them narrow (to be sharp and deep) or, the opposite, they were too broad and too long (in case of books), or they were written a while ago and may need an update. All these reasons motivated this work.

Thus, this monograph aims to complement and update the existing literature, aiming to be fairly broad, yet with some rigor, and yet also be relatively concise, with numerous references where more details can be found.

Along the way I realized the goal was too ambitious for one monograph and so I had to split it into two parts. In this first part that you are reading now I focus on the economic theory foundation for the Efficiency and Productivity Analysis. In the second part I plan to focus on the Data Envelopment Analysis (DEA). Again, the goal is to be concise and so, rather than trying to 'cover everything' (which is impossible), I try to give a foundation that an interested reader may find useful before learning more, whether it be in terms of theory or in terms of empirical work involving any estimator, and whether it is the DEA or its alternatives, such as Stochastic Frontier Analysis (SFA) or a symbiosis of them. This (hopefully) makes this part useful by itself, even if the reader decides not to pursue with DEA, SFA or any other particular approach.

This book can be also viewed as a much shorter (yet still fairly rigorous) discussion as those in Chapter 1-5 and 7 of Sickles and Zelenyuk (2019) and even shorter than the classic book of Färe and Primont (1995). It is also cast at a different angle than these sources which, *inter alia*, I try to follow myself and often refer to here for further details.

Chapter 1

Profit and Efficiency

1.1 General Profit Maximization

What is the main goal of a firm? The key premise of the mainstream economics is that a firm aims to *maximize its profit.* If one accepts this premise and wants to measure the performance of a firm, then the ultimate performance measure (at least within the neoclassical economics paradigm) should be based on profit, i.e., a profit efficiency measure. To state this more formally, and to facilitate further formal discussions, we will denote inputs and outputs as column vectors of dimensions N and M, namely $x = (x_1, ..., x_N)' \in \Re^N_+$ and $y = (y_1, ..., y_M)' \in \Re^M_+$, respectively.[1] In gen-

[1]It is worth noting that, theoretically, this framework allows for any finite dimension of inputs and outputs, however large they are, as long as they are finite dimensions. In particular, this theoretical framework also allows for the so-called big wide data contexts where N and M can be very large, even larger than the sample size, although practical implementation of it would require estimators that are adequately adapted to such contexts (e.g., see discussions in Zelenyuk (2020b)). In principle, it is also possible to cast this theory

eral, the profit maximization problem of a firm can be formally stated as

$$Maximal\ Profit := \max_{x,y}\{p(y)y - w(x)x \ : \ x \text{ can produce } y\}, \tag{1.1}$$

i.e., a firm will search for a technologically feasible optimal allocation (x, y) that will give the highest profit that it can achieve given its possibilities to use the vector of inputs x, purchased at some corresponding vector of input prices $w(x) \in \Re^N_+$, so that the firm can produce the vector of outputs y, and sell them at some corresponding vector of output prices $p(y) \in \Re^M_+$. Note that in this general formulation, output prices may depend on how much of y the firm wants to produce and sell, while the input prices may depend on how much of x the firm wants to buy.[2]

While stated concisely in (1.1), it should be clear that, in general, such a profit optimization problem is a very complex problem to solve in practice. Indeed, in practice, firms typically operate in dynamic and uncertain environments where the prices, technology and information that the economic agents of interest have about them at a given point of time are changing over time. As a result, such decisions might have to be taken in stages and with updates and adjustments with the arrival of new information. Here, we will mainly focus on the static case and so we do not have the subscript t for

in a more general framework allowing for firms owning different sub-firms (departments or plants) with potentially different dimensions of inputs and outputs in each sub-firm, by embedding the problem into the Hilbert sequence spaces, as in Färe and Zelenyuk (2021b).

[2] Obviously, the phrase 'x can produce y' may imply different meanings for different firms and at different times and is a generic statement that can be replaced with a more formal statement as we will do in the next section.

most of the discussion, except for Chapter 3 and parts of Chapter 4 where we also consider some aspects of the dynamics.

For a long time, a large part of economics literature has been assuming that firms (or individuals) somehow are able to achieve the goal of maximizing profit (or utility, for an individual). In some sense, such a 'belief' about full efficiency relieves from worrying about or dealing with possible inefficiency. This is indeed how many economics models have been developed, to simplify the matters. The reality, of course, appears to be very different. Indeed, many people (especially non-economists) agree such an idealistic situation does not always happen or perhaps even stronger: hardly ever happens in practice. In recent decades, this issue has been challenged and studied in behavioral economics, e.g., Thaler and Sunstein (2009, p. 6) eloquently noted the following in their famous book:

> "Individuals make pretty bad decisions in many cases because they do not pay full attention in their decision-making (they make intuitive choices based on heuristics), they don't have self-control, they are lacking in full information, and they suffer from limited cognitive abilities."

All these realistic (and perhaps natural for humans) phenomena lead to inefficiency, sometimes substantial, and sometimes perhaps causing or amplifying an economic crisis. Hence, while one may continue pretending there is no inefficiency or that it is not important, one may also (or instead) recognize that inefficiency may exist and, perhaps, may be a source of many societal problems. Such recognition will encourage researchers to search for the reasons of inefficiency, removal of which may improve the ways society operates and thus, hopefully, help improve its prosperity.[3]

Over the last few decades, more and more economists have been joining the rest of the people admitting an existence of inefficiency and thus understanding the need to look into the question of

(i) appropriate measurement and estimation of inefficiency;

(ii) analyzing and understanding the reasons or sources for the inefficiency, to help to find ways to remove or mitigate these reasons, in the hope of improving the efficiency of an economic system of interest.

In a sense, it is like in a sports race, say running a certain distance, e.g., a marathon: *A priori* everyone tries their best and is (perhaps) efficient in some respects, yet some happen to perform better than others and hence it is useful to measure these differences in performance and understand the factors associated with them, to understand how to improve the performance of an individual or a group of them.

The rest of this monograph is dedicated to the question of measurement, because indeed, as they say in management sciences (usually attributing it to Peter Drucker):

[3] Also see the influential works of Leibenstein (1966, 1975, 1987); Leibenstein and Maital (1992) for related discussions.

"If we can't measure it we can't manage it!"

The focus of this monograph is on the theoretical or conceptual measurement side that provides the foundation for statistical modeling, estimations and inference. Importantly, to outline this foundation, we will still deploy, with some adaptations, the rich mainstream microeconomics theory toolbox (involving profit functions, revenue functions, cost functions, technology characterizations, optimization, etc.).[4]

1.2 The Highest Efficiency Benchmark

To answer the question of how to measure efficiency, one first needs to identify what is an appropriate (and perhaps the best for the goals or circumstances) benchmark to measure an efficiency in general and the profit efficiency or its special cases in particular. From a social point of view, such a benchmark can be provided by the 'perfect competition' framework. *Inter alia*, this framework implies that the prices for all inputs and all outputs are exogenous to any individual firm, in the sense that individually any firm cannot influence them.[5] Thus, with such 'exogeneity of prices', the optimization criterion (1.1)

[4] While we will try to be self-contained and intuitive, a more advanced reader may refer to Mas-Colell et al. (1995) or Rubinstein (2006) for a thorough treatment of these concepts.

[5] E.g., see Mas-Colell et al. (1995) or Rubinstein (2006) for a textbook discussion of this framework and all the assumptions involved.

can be re-stated in terms of the *profit function*, defined as

$$\mathcal{PF}(p, w|\Psi \cap \mathcal{Z}) = \sup_{x,y}\{py - wx \; : \; (x, y) \in \Psi \cap \mathcal{Z}\}, \tag{1.2}$$

where, note, for an individual firm, p is no longer a function of y, while w is no longer a function of x. We also replaced the verbal statement of production feasibility, stated as 'x can produce y', with a more formal statement of the same meaning: $(x, y) \in \Psi$, where Ψ is the set of all production possibilities or input-output allocations (x, y), feasible for the firm at a moment of interest. The latter is often referred to as the *technology set*, defined in general terms as

$$\Psi = \{(x, y) \; : \; x \in \Re^N_+ \text{ can produce } y \in \Re^M_+\} \tag{1.3}$$

with some standard regularity conditions or axioms imposed on this set, as we detail below.

Furthermore, $\mathcal{Z}$ is a set of constraints relevant to the analyzed firm or its business activities, which could be permanent or temporary. This set of constraints may also serve to regularize scenarios where profit maximization may yield profit being 0 or ∞, as in the case of constant or increasing returns to scale, which we will discuss in more details later on. For example, such constraints may include limits on some inputs or outputs. Some of these constraints can be the reasons for the inefficiency a particular firm is experiencing. For simplicity of notation, we will assume that $\Psi = \Psi \cap \mathcal{Z}$ and is such that the maxima in (1.1) exist and so we can drop $\mathcal{Z}$ from our notation, unless explicitly needed otherwise.[6]

1.3 Assumptions about Technology

Many assumptions can be made about the technology set Ψ and in this sections we will consider the most commonly used in the literature.

1.3.1 Key Axioms about Technology

The axioms for technology set Ψ vary depending on the context and goals of modeling. Typically, they include the following properties:[7]

A1. Ψ is closed.

A2. The *output set* associated with the technology Ψ, defined as

$$P(x) := \{y \in \Re_+^M : (x, y) \in \Psi\} \tag{1.4}$$

is bounded $\forall x \in \Re_+^N$.

[6] Also see classic works of Diewert (1973); Lau (1976); McFadden (1978) about profit functions.

[7] See Shephard (1953, 1970) and Sickles and Zelenyuk (2019, Chapter 1) for more details and references.

A3. No 'free lunch', i.e.,[8]

$$(0_N, y) \notin \Psi, \;\; \forall y \geq 0_M. \tag{1.5}$$

A4. Producing nothing is possible, i.e.,

$$0_M \in P(x), \; \forall x \in \Re^N_+. \tag{1.6}$$

A5a. Free disposability (also called 'strong disposability') of outputs, i.e.,

$$\forall (x_o, y_o) \in \Psi \Rightarrow (x_o, y) \in \Psi, \;\; \forall y \leqq y_o. \tag{1.7}$$

A5b. Free disposability (also called 'strong disposability') of inputs, i.e.,

$$\forall (x_o, y_o) \in \Psi \Rightarrow (x, y_o) \in \Psi, \quad \forall x \geqq x_o. \tag{1.8}$$

These axioms help regularizing the modeling, and hence often referred to as regularity conditions. E.g, they help ensuring existence of optima, since A1 and A2 imply $P(x)$ is compact as is the relevant portion of Ψ (possibly with additional constraints) with respect to which the optimization of

[8]Note on the notation for vectors we will use throughout: for any $a, z \in \Re^M$ (i.e., vectors a, z of a dimension $M \times 1$), "$a > z$" or "$z < a$" $\iff a - z \in \Re^M_{++}$, i.e., all elements of a are greater than the corresponding elements of z; "$a \geqq z$" or "$z \leqq a$" $\iff a - z \in \Re^M_+$, i.e., all elements of a are larger than or equal to the corresponding elements of z; "$a \geq z$" or "$z \leq a$" $\iff a - z \in \Re^M_+ \backslash \{0_M\}$, where 0_M is the vector where all elements are zeros (i.e., "$a \geqq z$" or "$z \leqq a$" *and* at least one element of a is greater than the corresponding element of z).

profit is made. Axiom A3 and A4 imply $(0_N, 0_M) \in \Psi$, which is important for some proofs in economic theory, although these 'singularity' points are typically assumed away for the applications (e.g., to avoid dividing by zero). Together, A1-A4 usually constitute the minimal set of conditions imposed on technology. Axiom A5 (a and b) basically imply monotonicity of the technology process in the sense that increase in inputs cannot lead to decrease of outputs. These are also most commonly assumed, usually for convenience (or perhaps due to ignorance) even though they are often not entirely needed and can be replaced with a weaker version, as we discuss in the next section.

1.3.2 Weak disposability

Some contexts may require acknowledging that the disposability is not 'free' ('strong'), and modify it to a 'weaker' version. For example, some production processes may require disposing of (reducing) some outputs jointly (e.g. in some proportion) with other specific outputs. This phenomenon is typical for production processes involving some 'undesirable' or 'bad' outputs, such as pollution (e.g. CO_2, etc.). In this case, A5a can be modified as following:

A5aw. Weak disposability of outputs, i.e.,

$$\forall (x_o, y_o) \in \Psi \Rightarrow (x_o, \theta y_o) \in \Psi, \quad \forall \theta \in [0, 1]. \tag{1.9}$$

Similarly, some production processes may be prone to congestion phenomenon, i.e., when the increase of some inputs may actually cause reduction in outputs, unless some other inputs are also increased.

A5bw. Weak disposability of inputs, i.e.,

$$\forall (x_o, y_o) \in \Psi \Rightarrow (x_o/\theta, y_o) \in \Psi, \quad \forall \theta \in (0, 1]. \tag{1.10}$$

It is also possible to impose weak disposability only on some outputs (or inputs), while keeping it free for the other outputs (or inputs). For example, we can partition vector y into (g, b), where g is a vector of good outputs for which free disposability applies and b is a vector of 'bad' outputs for which only weak disposability (jointly with good outputs) applies.

A5a1. Free (strong) disposability of desirable outputs:

$$\forall (x_o, g_o, b_o) \in \Psi \Rightarrow (x_o, g, b_o) \in \Psi, \quad \forall g \leqq g_o.$$

A5a2. Jointly weak disposability of desirable outputs and undesirable outputs:

$$\forall (x_o, g_o, b_o) \in \Psi \Rightarrow (x_o, \theta g_o, \theta b_o) \in \Psi, \quad \forall \theta \in [0, 1].$$

More discussions on this can be found in Shephard (1974), Färe and Svensson (1980), Färe and Grosskopf (1983), Färe and Grosskopf (2009), Podinovski and Kuosmanen (2011), Pham and Zelenyuk (2019) and references therein.

1.3.3 Returns to Scale and Homotheticity

While in practice such an idealistic framework as perfect competition rarely occurs, it still serves as a valuable benchmark for measuring performance (without assuming the achievement of perfect competition), at least from the perspective of a society that strives to fully utilize its resources. Thus, considering it here does not mean we assume its occurrence, rather we identify the ideals to strive and compare to. The profit maximization under perfect competition in the long-run is also cohesive with the notion of the *constant returns to scale* (CRS) property of technology. This property can be defined in the *global* and *local* senses. In the global sense, Ψ exhibits CRS if and only if

$$\delta\Psi = \Psi, \ \forall\delta > 0, \tag{1.11}$$

i.e., if and only if technology set Ψ is a cone. Intuitively, it means that whenever technology is such that any equiproportional change (by any scaling factor $\delta > 0$) in all inputs leads to the same equiproportional change (by the same factor $\delta > 0$) in all the outputs. This notion is somewhat idealistic, and also hardly ever happens in practice in its global sense (i.e., $\forall\delta > 0$), yet it serves as a useful benchmark for measuring what a society might strive to achieve, at least locally and hypothetically. Thus, we will frequently encounter this notion in the later discussions.

While being useful as the highest benchmark to measure with respect to, perfect competition (or the CRS) might also be viewed as 'too demanding' for a criterion to accept in some or even many cases, e.g., in a short run when firms might be less flexible for various reasons that are specific to particular regions, times, industries or firms. So, it is useful to consider other appropriate criteria. Indeed, consider a case where inputs are less flexible than outputs (or even must be fixed for a period of time), e.g., the contracts for buildings, labour and capital have been signed for the upcoming year and are very costly to break. In this case, one could think that a natural benchmark would be the maximal revenue rather than maximal profit. On the other hand, if the outputs are less flexible (e.g., if the marketing department or a regulator has defined certain levels of output to be produced, with little flexibility) and the firm needs to optimize their inputs to produce those levels of output, then the cost minimization criterion would be more natural. We will consider these cases in more details in the next sections, starting with the general profit efficiency and then move to more specific efficiency measures.

Meanwhile, it is worth noting that one can impose many other properties on technology. For example, Ψ exhibits *global non-increasing returns to scale* (NIRS) if and only if

$$\delta\Psi \subseteqq \Psi, \forall\delta \in (0,1), \tag{1.12}$$

while Ψ exhibits *global non-decreasing returns to scale* (NDRS) if and only if

$$\delta\Psi \subseteqq \Psi, \forall\delta \in (1,+\infty). \tag{1.13}$$

1.3.4 Separability

Different types of separability are also sometimes imposed as structural assumptions on technology. The most common of these is the so-called *homotheticity,* which extends the notion of constant returns to scale (e.g., see Shephard (1953, 1970)). For example, technology is called *output homothetic* if and only if we have the following decomposition or separability of the output sets in the sense that

$$P(x) = G(x)P(1_N), \quad \forall x \in \Re^N_+, \tag{1.14}$$

for some suitable function $G(x)$ coherent with the regularity conditions on Ψ.

Similarly, the technology is called *input homothetic* if and only if we have the following decomposition or separability on the input side

$$L(y) = H(y)L(1_M), \quad \forall y \in \Re^M_+, \tag{1.15}$$

where $H(y)$ is a suitable function coherent with the regularity conditions on Ψ and where $L(y)$ is the *input requirement set* associated with Ψ, defined as

$$L(y) = \{x \in \Re^N_+ : (x, y) \in \Psi\}, \quad y \in \Re^M_+. \tag{1.16}$$

Furthermore, whenever technology exhibits both input homotheticity and output homotheticity, it is referred to as *jointly input and output homothetic* or *inversely homothetic*.[9]

Finally, when considering the dynamic aspects, one sometimes also imposes various structures on how the technology is changing over time, e.g., the so-called Hicks-neutral technology change or 'Hicks-neutrality', which we will consider in more detail in Chapter 3.

1.3.5 Convexity and Quasi-concavity

In some contexts, researchers may want to assume some variants of convexity. For example, to achieve some of the duality theory results about the profit function one needs to assume that Ψ is convex, i.e.,

$$\forall (x_0, y_0),\ (x_1, y_1) \in \Psi \Rightarrow \theta(x_0, y_0) + (1-\theta)(x_1, y_1) \in \Psi,\ \forall \theta \in [0,1]. \quad (1.17)$$

A much weaker requirement is to assume that the output sets are convex, i.e.,

$$\forall y_0,\ y_1 \in P(x) \Rightarrow \theta y_0 + (1-\theta) y_1 \in P(x),\ \forall x \in \Re^N_+, \forall \theta \in [0,1], \quad (1.18)$$

which is required to achieve various results in duality theory about the revenue function. Another (weaker than (1.17)) requirement is to assume that the input sets are convex, i.e.,

$$\forall x_0,\ x_1 \in L(y) \Rightarrow \theta x_0 + (1-\theta) x_1 \in L(y), \quad \forall \theta \in [0,1], \forall y \in \Re^M_+, \quad (1.19)$$

[9] Also see Hanoch (1970); Färe and Mitchell (1993); Chambers and Färe (1998); Färe et al. (2001); Zelenyuk (2014b); Färe et al. (2020) for further discussions.

which is required to achieve some of the duality theory results about the cost function.[10]

It is also worth noting that convexity of Ψ implies convexity of $P(x)$ and of $L(y)$, for all $y \in \Re_+^M$ and all $x \in \Re_+^N$, while convexity of $P(x)$ or $L(y)$ does not imply convexity of Ψ. Moreover, convexity of $L(y)$ does not imply and is not implied by the convexity of $P(x)$.

Also note that convexity of $P(x)$ is equivalent to saying that the correspondence that generates $L(y)$ is quasi-concave. Similarly, convexity of $L(y)$ is equivalent to saying that the correspondence that generates $P(x)$ is quasi-concave (e.g., see Pham and Zelenyuk (2019)).

1.4 General Profit Efficiency

A myriad of efficiency measures were proposed in the literature spanning a few decades, which also included many profit efficiency measures. A recent review, with an attempt for unification of many of these measures, was endeavored by Färe et al. (2019) with some further developments in Färe and Zelenyuk (2020). Following these works, we start with a very general measure of *profit efficiency*, defined, for an (observed or hypothetical) input-output

[10]For details on duality in economics, see Shephard (1953, 1970), Diewert (1971, 1974, 1982) and a brief textbook style discussion in Chapter 2 of Sickles and Zelenyuk (2019).

allocation (x^j, y^j) and prices (w, p), as following

$$\begin{aligned} \mathcal{GPE}(x^j, y^j; w, p|\Psi) = \sup_{\gamma,\lambda,x,y} \{\psi(\gamma, \lambda) :& \\ & \psi_y(p, y^j, \gamma) - \psi_x(w, x^j, \lambda) \leq py - wx, \\ & (x, y) \in \Psi,\ \gamma \in \Gamma,\ \lambda \in \Lambda\}, \end{aligned} \tag{1.20}$$

where $\psi(\gamma, \lambda)$ is an objective function selected by a researcher, to be optimized over $\gamma = (\gamma_1, \dots, \gamma_M)$ and $\lambda = (\lambda_1, \dots, \lambda_N)$ as well as (x, y), while $\psi_x(w, x, \lambda)$ and $\psi_y(p, y, \gamma)$ are the functions that specify the way the measurement of the efficiency must be done with respect to each element of input and output vectors, and where Γ and Λ are the sets specifying permissible values of γ and λ, respectively. For example, and as in Färe et al. (2019), one may specify $\psi_y(p, y^j, \gamma) = \sum_{m=1}^{M} p_m \gamma_m y_m^j$ and $\psi_x(w, x^j, \lambda) = \sum_{l=1}^{N} w_l \lambda_l x_l^j$, with $\Gamma = \Re_+^M$ and $\Lambda = \Re_+^N$.[11]

As was pointed out by Färe and Zelenyuk (2020), the optimization problem (1.20) is a way to formalize a general modeling framework with 'producer preferences' embedded in an optimization criterion that is more general than a profit function, with the latter as a special case, allowing to incorporate various objectives of an economic agent. Moreover, it can also be viewed as a dual counterpart of a general technical efficiency measure that satisfies the

[11] Note that here we use superscript j to distinguish the particular allocation of interest (x^j, y^j) for which the measurement of efficiency is performed from a generic allocation (x, y) which is optimized over in (1.20). Whenever this distinction is not important we will be dropping j for simplicity.

so-called *Pareto-Koopmans efficiency criterion.* More formally, this criterion defines the so-called *Pareto-Koopmans efficient subset of* Ψ*,* defined as

$$\mathit{eff}\,\partial\Psi \equiv \{(x,y)\ :\ (x,y)\in\Psi,\ (x^0,y^0)\notin\Psi, \forall(-x^0,y^0)\geq(-x,y)\}. \quad (1.21)$$

That is, an allocation (x, y) is *Pareto-Koopmans efficient* for technology Ψ if and only if with such technology it is impossible to produce more of any outputs without decreasing some other outputs or increasing some of the inputs, and it is also infeasible to reduce any of the inputs without increasing some other inputs or decreasing some of the outputs. It should be also clear that the efficiency criterion (1.20) is superior relative to the Pareto-Koopmans efficiency, because it is taking into account the efficiency with respect to technology as well as the prices that reflect the current valuations in the markets for each input and each output.

Importantly, many efficiency measures suggested in the literature can be derived from (1.20), including the *Nerlovian (or directional) profit efficiency measure* (Chambers et al. (1996, 1998)), defined as

$$\mathcal{NPE}_d(x^j, y^j, p, w|\Psi) \equiv \frac{\mathcal{PF}(w,p|\Psi) - (py^j - wx^j)}{(pd_y + wd_x)}, \quad (1.22)$$

where $d = (-d_x, d_y)$ is a nonzero vector in $\Re^N_- \times \Re^M_+$ that specifies the direction in which the distance between (x^j, y^j) and the frontier of Ψ is to be measured, and provided that $pd_y + wd_x \neq 0$.[12] Indeed, and as pointed out by

[12] It is worth clarifying the common notation we use here: the subscripts in d_x and

Färe and Zelenyuk (2020), (1.22) is obtained from (1.20) when $\psi(\gamma, \lambda) = \alpha$, such that $\gamma = \alpha d_y$ and $\lambda = \alpha d_x$ and let $\psi_y(p, y^j, \gamma) = p(y^j + \alpha d_y)$ and $\psi_x(w, x^j, \lambda) = w(x^j - \alpha d_x)$.[13]

1.5 Output and Input Oriented Profit Efficiency

Nerlovian profit efficiency measure is just one of many interesting special cases that can be obtained from (1.20). Another interesting case is derived by letting $\lambda_1 = \lambda_2 = \ldots = \lambda_N = 1$ and $\gamma_1 = \gamma_2 = ... = \gamma_M = \gamma$ and $\psi(\lambda_1, \ldots, \lambda_N; \gamma_1, \ldots, \gamma_M) = \gamma$ and also setting $\psi_y(p, y^j, \gamma) = \sum_{m=1}^{M} p_m \gamma_m y_m^j$ while $\psi_x(w, x^j, \lambda) = \sum_{l=1}^{N} w_l \lambda_l x_l^j$ and thus leading to the *output-oriented profit efficiency measure* and assuming $py^j \neq 0$, given by

$$\mathcal{OPE}(x^j, y^j; w, p|\Psi) = \sup_{\gamma} \left\{ \sup_{x,y} \left\{ \frac{py - wx + wx^j}{py^j} \ : \ (x, y) \in \Psi \right\} \geq \gamma \right\}. \tag{1.23}$$

Proposing this measure, Färe et al. (2019) also gave several intuitive interpretations through simple notions often used in business analysis. For

d_y indicate that those are vectors in the same spaces as x and y, i.e., in $\Re_+^N$ and $\Re_+^M$, respectively, and may or may not depend on x or y. For example, as we will see later on, the popular choices of $(-d_x, d_y)$ in the literature include: $(-1_N, 1_M)$, $(0_N, 1_M)$, $(-1_N, 0_M)$, $(-x, y)$, $(-x, 0_M)$, $(0_N, y)$.

[13] Also see related discussions in Kumbhakar (1987); Foster et al. (2008); Cooper et al. (2011a); Aparicio et al. (2013); Färe et al. (2015), to mention a few.

example, we have

$$\begin{aligned}\mathcal{OPE}(x^j, y^j; w, p|\Psi) =& 1 + \frac{\mathcal{PF}(p, w|\Psi) - \pi^j}{r^j} \\ =& \frac{c^j}{r^j} + \frac{\mathcal{PF}(p, w|\Psi)}{r^j},\end{aligned} \tag{1.24}$$

where $c^j = wx^j$, $r^j = py^j$ and $\pi^j = r^j - c^j$ are the observed total costs, observed total revenue and observed profit, respectively, for the allocation (x^j, y^j) and prices (w, p).

That is, the output oriented profit efficiency measure (1.24) or (1.23) can be decomposed into two parts, each of which represents a key performance indicator for a business analysis. The first component is the realized cost-benefit ratio, which is also the reciprocal of what Georgescu-Roegen (1951) called the '*return to the dollar*' measure of performance. The second component is the best possible *profit margin* for the allocation (x^j, y^j) with prices (w, p).

It is also worth recalling here that the idealistic framework of the perfect competition implies that the long-run equilibrium is achieved when the economic profit is zero. In the light of expression (1.24), it means that all that matters for the performance from the perspective of long-run perfectly competitive equilibrium is the first component—the cost-benefit ratio, which is indeed one of the most important *key performance indicators* (KPIs) in business, finance and economic analysis in practice.

Another interesting case from Färe et al. (2019) is derived when $\psi_y(p, y^j, \gamma) =$

$\sum_{m=1}^{M} p_m \gamma_m y_m^j$ and $\psi_x(w, x^j, \lambda) = \sum_{l=1}^{N} w_l \lambda_l x_l^j$ and we let $\lambda_1 = \lambda_2 = \ldots = \lambda_N = \lambda$ and $\gamma_1 = \gamma_2 = ... = \gamma_M = 1$ and $\psi(\lambda_1, \ldots, \lambda_N; \gamma_1, \ldots, \gamma_M) = 2 - \lambda$, which gives an *input-oriented profit efficiency measure*, given by

$$\mathcal{IPE}(x^j, y^j; w, p|\Psi) = 1 + \frac{\mathcal{PF}(p, w|\Psi) - \pi^j}{c^j} \tag{1.25}$$

assuming that $wx^j \neq 0$.

Chapter 2

Cost, Revenue and Technical Efficiency

While the maximal profit can be viewed as an ultimate benchmark for a profit-maximizing entity, various circumstances of reality may force such an entity (or a researcher analyzing it) to pursue other alternatives, such as Cost Efficiency, Revenue Efficiency and Technical Efficiency, which we will briefly cover in this chapter.

2.1 Three in One: Profit, Cost and Revenue Efficiency

Following Färe and Zelenyuk (2020), let us consider another useful decomposition that has the main trinity of efficiency analysis—profit efficiency, cost efficiency and revenue efficiency—in one expression. Indeed, in the example of output-oriented profit efficiency, (1.23), we have:

$$\begin{aligned} \mathcal{OPE}(x^j, y^j; w, p|\Psi) &= \frac{\mathcal{PF}(p, w|\Psi)}{r^j} + \\ + \, \mathcal{CE}(x^j, y^j, w|\Psi) &\times \mathcal{RE}(x^j, y^j, p|\Psi) \times \mathcal{CBR}^*(x^j, y^j, w, p|\Psi), \end{aligned} \tag{2.1}$$

where $\mathcal{CE}(x^j, y^j, w|\Psi)$ is a measure of *cost efficiency* for the same allocation (x^j, y^j), input prices w and technology Ψ, defined as

$$\mathcal{CE}(x^j, y^j, w|\Psi) = \frac{wx^j}{\mathcal{C}(y^j, w|\Psi)}, \tag{2.2}$$

i.e., in words, the observed cost divided by the optimal cost given by the classical *cost function*, $\mathcal{C}(y, w|\Psi)$, defined as

$$\mathcal{C}(y, w|\Psi) = \inf_x \{wx \; : \; (x, y) \in \Psi\}, \tag{2.3}$$

while $\mathcal{RE}(x^j, y^j, p|\Psi)$ is a measure of *revenue efficiency* for allocation (x^j, y^j), output prices p and technology Ψ, defined as

$$\mathcal{RE}(x^j, y^j, p|\Psi) = \frac{\mathcal{R}(x^j, p|\Psi)}{py^j}, \tag{2.4}$$

where $\mathcal{R}(x, p|\Psi)$ is the classical *revenue function* defined as

$$\mathcal{R}(x, p|\Psi) = \sup_y \{py \; : \; (x, y) \in \Psi\}, \tag{2.5}$$

and, finally, $\mathcal{CBR}^*(x^j, y^j, w, p|\Psi)$ is the cost-benefit ratio that compares the optimal cost (from the cost-minimization perspective (2.3)) to optimal revenue (from the revenue-maximization perspective (2.5)), i.e.,

$$\mathcal{CBR}^*(x^j, y^j, w, p|\Psi) = \frac{\mathcal{C}(y^j, w|\Psi)}{\mathcal{R}(x^j, p|\Psi)}. \tag{2.6}$$

In their own turns, the cost efficiency and the revenue efficiency can be further decomposed into technical efficiency and allocative efficiency, as we discuss in the next section. Analogous decompositions can be made for other orientations and we leave it for the readers.

2.2 Technical Efficiency and Duality

When price information is not available, or not reliable, a researcher may resort to employing the notion of technical efficiency—a gap between an input-output allocation (x, y) and the boundary of the technology set Ψ. Because there are many ways to measure this gap (e.g., along many different directions), there are many alternatives of the technical efficiency and here we will consider a few of the most popular ones in practice.

2.2.1 Directional Distance Function

One of the most general measures of technical (in)efficiency is the so-called *directional distance function,* $\mathcal{DDF}_d : \Re^N_+ \times \Re^M_+ \to \Re \cup \{+\infty\}$, defined as[1]

$$\mathcal{DDF}_d(x, y|\Psi) \equiv \sup_\beta \{\beta \in \Re \ : \ (x, y) + \beta d \in \Psi\}, \tag{2.7}$$

which measures the distance between any point $(x, y) \in \Re^N_+ \times \Re^M_+$ and the boundary of Ψ in terms of a portion of a pre-specified non-zero vector $d = (-d_x, d_y) \in \Re^N_- \times \Re^M_+$ that also defines the direction (or orientation) of the measurement.

This function gives a *complete characterization of technology* Ψ, in the sense that for any $d = (-d_x, d_y) \in \Re^N_- \times \Re^M_+ \backslash \{0_{N+M}\}$, we have

$$\mathcal{DDF}_d(x, y|\Psi) \geq 0 \iff (x, y) \in \Psi. \tag{2.8}$$

Moreover, this function can be (and often is) used as a measure of technical (in)efficiency, as it can indicate that (x, y) is technically efficient if and only if $\mathcal{DDF}_d(x, y|\Psi) = 0$ and is inefficient otherwise. The size of inefficiency is therefore given by the value of $\mathcal{DDF}_d(x, y|\Psi)$ and, clearly, depends on the chosen directional vector, as it represents the maximal value of that vector that can be added to (x, y) to project it to the boundary of Ψ.

Importantly, note that there is also a duality between the DDF and the

[1]See Chambers et al. (1996, 1998) and Sickles and Zelenyuk (2019) for more details.

profit function (or profit efficiency), since the latter can give a dual characterization of the use of technology that also incorporates price information. While the related duality theory is a subject in itself, it can be concisely summarized via the following two expressions (if we assume that Ψ is convex, and for any $d = (-d_x, d_y) \in \Re^N_- \times \Re^M_+ \backslash \{0_{N+M}\}$ and $pd_y + wd_x \neq 0$:

$$\mathcal{PF}(w,p) = \sup_{x,y}\{py - wx + \mathcal{DDF}_d(x,y|\Psi)(pd_y + wd_x)\},\ (w,p) \in \Re^{N+M}_{++} \quad (2.9)$$

if and only if

$$\mathcal{DDF}_d(x,y|\Psi) = \inf_{w,p}\left\{\mathcal{NPE}_d(x,y,p,w|\Psi)\right\},\ (x,y) \in \Re^{N+M}_{+}. \quad (2.10)$$

2.2.2 Shephard's Distance Functions and Debreu-Farrell Measures

Other popular technical efficiency measures can be derived from $\mathcal{DDF}_d(x,y|\Psi)$. For example, by setting $d = (0, y)$, we get

$$\mathcal{DDF}_d(x,y|\Psi) = \frac{1}{\mathcal{ODF}(x,y|\Psi)} - 1, \quad (2.11)$$

where $\mathcal{ODF}(x,y|\Psi)$ stands for the *output oriented Shephard distance function*, $\mathcal{ODF} : \Re^N_+ \times \Re^M_+ \to \Re_+ \cup \{+\infty\}$, defined as

$$\mathcal{ODF}(x,y|\Psi) \equiv \inf_{\gamma}\{\gamma > 0 : (x, y/\gamma) \in \Psi\}, \quad (2.12)$$

which can in turn be used to define (and lay out the theory for) the *output oriented Debreu-Farrell measure of technical efficiency*, defined as the reciprocal of $\mathcal{ODF}(x, y|\Psi)$, i.e.,[2]

$$\mathcal{OTE}(x, y|\Psi) \equiv \frac{1}{\mathcal{ODF}(x, y|\Psi)}. \tag{2.13}$$

Intuitively, these functions tell us about the maximal possible expansion (projection) of the vector of all outputs, y, by some common scalar, γ, until it reaches the boundary of Ψ, at some point y^*, while keeping inputs x fixed.[3] These functions will indicate that (x, y) is technically efficient if and only if $\mathcal{OTE}(x, y|\Psi) = 1/\mathcal{ODF}(x, y|\Psi) = 1$ and otherwise, when $\mathcal{OTE}(x, y|\Psi) > 1$ (or $\mathcal{ODF}(x, y|\Psi) < 1$), is inefficient. Hence, the size of inefficiency is given by the value of $1 - \mathcal{ODF}(x, y|\Psi)$ when the efficient output is taken as the base for the comparison and so, the efficiency is measured on the scale $[0, 1]$. Alternatively, it can be measured by the value of $\mathcal{OTE}(x, y|\Psi) - 1$ when the actual output is taken as the base for the comparison and so, the efficiency is measured on the scale $[1, \infty]$.

[2]The idea of such measurements of efficiency appeared in Debreu (1951) and was more elaborated in Farrell (1957), hence the name, although its roots can be also found it von Neumann (1945).

[3]Geometrically, these functions can be represented as the ratios of the lengths of two vectors, y and y^* and the difference is which of these is placed at the base (i.e., in the denominator): $\mathcal{ODF}$ uses the efficient output for the base (i.e., $||y||/||y^*||$), while $\mathcal{OTE}$ uses the actual output for the base (i.e., $||y^*||/||y||$), which determines their intuitive interpretations.

Meanwhile, if we choose $d = (-x, 0)$, then we get

$$\mathcal{DDF}_d(x, y|\Psi) = 1 - \frac{1}{\mathcal{IDF}(x, y|\Psi)}, \tag{2.14}$$

where $\mathcal{IDF}(x, y|\Psi)$ stands for the *input oriented Shephard distance function*, $\mathcal{IDF} : \Re^N_+ \times \Re^M_+ \to \Re_+ \cup \{+\infty\}$, defined as

$$\mathcal{IDF}(x, y|\Psi) \equiv \sup_{\lambda}\{\lambda > 0 : (x/\lambda, y) \in \Psi\}, \tag{2.15}$$

which can in turn be used to define (and provide the theory for) the *input oriented Debreu-Farrell measure of technical efficiency*, defined as

$$\mathcal{ITE}(x, y|\Psi) \equiv \frac{1}{\mathcal{IDF}(x, y|\Psi)}. \tag{2.16}$$

Intuitively, $\mathcal{ITE}(x, y|\Psi)$ searches for the maximal possible contraction of all inputs x by some common scalar λ, so that it reaches (or is projected onto) the boundary of Ψ, at some point x^*, while keeping outputs y fixed.[4] These functions will indicate that (x, y) is technically efficient if and only if $\mathcal{ITE}(x, y|\Psi) = 1/\mathcal{IDF}(x, y|\Psi) = 1$ and otherwise, when $\mathcal{ITE}(x, y|\Psi) < 1$ (or $\mathcal{IDF}(x, y|\Psi) > 1$), is inefficient. Therefore, the size of inefficiency is given by the value of $1 - \mathcal{ITE}(x, y|\Psi)$ when the actual input is taken as the base and so, the efficiency measured on the scale $[0, 1]$, or by the value of $\mathcal{IDF}(x, y|\Psi) - 1$ when the efficient input is taken as the base and so, the

[4]Again, geometrically, these functions can be represented as the ratios of the lengths of two vectors, x and x^* and the difference is which of these is at the base (or denominator).

efficiency is measured on the scale $[1, \infty]$.

2.2.3 Properties of Distance Functions

Distance functions (directional and Shepard's) have been proven to possess useful mathematical properties that make them convenient tools in production analysis and economics in general. Among the most important properties they have is that of *complete characterization of technology* Ψ, which we have already stated for the directional distance function, in (2.8). Analogously, for the Shepard's distance functions the complete characterization properties are stated as

$$0 \leq \mathcal{ODF}(x, y|\Psi) \leq 1 \iff (x, y) \in \Psi \tag{2.17}$$

and

$$\mathcal{IDF}(x, y|\Psi) \geq 1 \iff (x, y) \in \Psi. \tag{2.18}$$

These, together with (2.8), imply that these distance functions are equivalent characterizations of the same technology, in the sense that

$$0 \leq \mathcal{ODF}(x, y|\Psi) \leq 1 \iff \mathcal{IDF}(x, y|\Psi) \geq 1 \iff \mathcal{DDF}_d(x, y|\Psi) \geq 0. \tag{2.19}$$

These functions also represent alternative (and equivalent in the sense of (2.19)) generalizations of the notion of the production function that is commonly used in economics. To be precise, recall that when there is only one output, then technology can be characterized via a production function, defined in general terms as

$$\mathcal{F}(x|\Psi) \equiv \max\{y : (x, y) \in \Psi\} \tag{2.20}$$

and so, after a bit of algebra (and using property (2.24) below), it follows that

$$\mathcal{ODF}(x, y|\Psi) = \frac{y}{\mathcal{F}(x|\Psi)}, \quad \forall (x, y) \in \Re^N_+ \times \Re^1_+ \tag{2.21}$$

i.e., in the case of single-output technologies, the $\mathcal{ODF}$ is simply the ratio of the actual output to maximal output, which is indeed a natural measure of efficiency in producing the single output y. Meanwhile, and as mentioned above, in the multi-output case, this ratio is generalized to be in terms of the ratio of the lengths of the vectors: $||y||/||y^*||$, where y^* is the radial projection of $y \in \Re^M_+$ onto the frontier of $P(x)$, while keeping x and technology the same.

Importantly, note that, in general, $\mathcal{ODF}(x, y|\Psi) = 1$ does not imply that (and is not implied by) $\mathcal{IDF}(x, y|\Psi) = 1$, neither do any of these imply (or are implied by) $\mathcal{DDF}_d(x, y|\Psi) = 0$. Thus, in general, different efficiency measures applied to the same allocation (x, y) may (and often do) yield different efficiency scores, sometimes very different. For example, one may suggest 100% efficiency while another may suggest high inefficiency. It is therefore very important to give careful consideration to and justifications for the choice of a technical efficiency measure or, possibly, use several alternatives to reach a robust conclusion.

It is also useful to know the conditions under which they yield the same or similar results. One such condition is the assumption of CRS, as defined in (1.11), which is theoretically equivalent to saying that

$$\mathcal{ITE}(x, y|\Psi) = \frac{1}{\mathcal{OTE}(x, y|\Psi)}, \ \forall (x, y) \in \Re^N_+ \times \Re^M_+. \tag{2.22}$$

A similar equivalence under CRS also exists for the relationships between the DDF and the Shephard's distance functions for a more general family of directional vectors than those considered above.[5]

Other useful properties include homogeneity (of degree 1) of the Shephard's distance functions: in x for the input oriented and in y for the output oriented.[6] That is, we have

$$\mathcal{IDF}(\delta x, y|\Psi) = \delta \mathcal{IDF}(x, y|\Psi), \quad \forall \delta > 0 \tag{2.23}$$

and

$$\mathcal{ODF}(x, \delta y|\Psi) = \delta \mathcal{ODF}(x, y|\Psi), \quad \forall \delta > 0. \tag{2.24}$$

Meanwhile, the analogous property for the DDF, is the property of 'translation by any real scalar', defined as

$$\mathcal{DDF}_d(x, y|\Psi) - \delta = \mathcal{DDF}_d(x - \delta d_x, y + \delta d_y|\Psi), \quad \forall \delta \in \Re.$$

[5]E.g. see Zelenyuk (2014a) and references therein for more details.

[6]In turn, this means that the corresponding Debreu-Farrell technical efficiency measures are homogeneous of degree -1.

For other useful properties of these functions (e.g., monotonicity, semi-continuity, semi-boundedness, etc.) see Russell (1990), Dmitruk and Koshevoy (1991), Färe and Primont (1995), Chambers et al. (1996) and more recently in a textbook style in Sickles and Zelenyuk (2019, Chapter 1 and 3) and references therein. For using derivative properties of these functions, e.g., in the context of calculating elasticities, see Zelenyuk (2013) and references therein.

2.2.4 Non-directional Measures of Technical Efficiency

The technical efficiency measures discussed above required a pre-specified direction of measurement. It is also possible to have efficiency measures that do not impose a particular direction to go along, and leave that for the optimization mechanism to figure it out. Such measures can also be deduced from (1.20). For example, among the most popular of such measures is the so-called (generalized) Russell-type measure of technical efficiency, which in general terms can be stated as

$$\begin{aligned} \mathcal{RTE}(x,y) \equiv \sup_{\gamma_1,...,\gamma_M,\lambda_1,...,\lambda_N} \Big\{ \psi(\gamma,\lambda) : \\ (x_1/\lambda_1,...,x_N/\lambda_N,\gamma_1 y_1,...,\gamma_M y_M) \in \Psi, \\ \lambda_1,...,\lambda_N,\gamma_1,...,\gamma_M \geq 1 \Big\}, \quad (2.25) \end{aligned}$$

where $\psi(\gamma, \lambda)$ is an appropriate aggregator of the individual scaling factors. For example, one could choose an additive aggregator for ψ, i.e.,

$$\psi(\gamma, \lambda) \quad = \quad \left(\sum_{m=1}^{M} \mathbb{I}_0(y_m)\right)^{-1} \sum_{m=1}^{M} \gamma_m \mathbb{I}_0(y_m) + \left(\sum_{l=1}^{N} \mathbb{I}_0(x_l)\right)^{-1} \sum_{l=1}^{N} \lambda_l \mathbb{I}_0(x_l)$$

where $\mathbb{I}_0(A) = 1$ if $A > 0$ and 0 otherwise, which plays the role of a cancelling device, to remove the expansion and contraction scaling factors from the inputs or outputs that are zero. A special case of ψ, for the input orientation, was originally considered by Färe and Lovell (1978) who proposed and named this measure, while its output oriented analogue can be found in Sickles and Zelenyuk (2019).[7]

Furthermore, it is also possible to deduce the *generalized hyperbolic measure* of technical efficiency, defined as

$$\mathcal{HTE}(x, y) \equiv \sup_{\lambda,\gamma} \left\{\psi(\gamma, \lambda) : (x/\lambda, \gamma y) \in \Psi, \ \lambda, \gamma \geq 1\right\}, \tag{2.26}$$

where $\psi(\gamma, \lambda)$ is an appropriate aggregator of the individual scaling factors, e.g., $\psi(\gamma, \lambda) = (\gamma + \lambda)/2$ or $\psi(\gamma, \lambda) = \sqrt{\gamma \times \lambda}$ or their weighted versions. That is, this measure expands all outputs by a common scalar γ, while at the same time contracting all inputs by a common scalar λ, until the allocation $(x/\lambda, \gamma y)$ reaches the frontier. Thus, the shape of the trajectory of the projection of a point (x, y) to the frontier reminds the shape of the hyper-

[7] One may also use a multiplicative alternative for $\psi(\gamma, \lambda)$, e.g. as in Färe et al. (2007).

bola, hence its name. The exact trajectory depends on $\psi(\gamma,\lambda)$. A special (simplified and most popular) case of this measure is when $\psi(\gamma,\lambda)=\gamma=\lambda$.

Finally, it is also possible to deduce the so-called *slack-based (or additive) measures* of technical efficiency, defined in general terms as

$$\begin{aligned}\mathcal{SBTE}(x,y) \equiv \sup_{s_1^-,\ldots,s_N^-,s_1^+,\ldots,s_M^+} \Big\{ & \psi(s_1^-,...,s_N^-,s_1^+,...,s_M^+) : \\ & (x_1-s_1^-,...,x_N-s_N^-,y_1+s_1^+,...,y_M+s_M^+)\in\Psi, \\ & s_1^-,...,s_N^-,s_1^+,...,s_M^+ \geq 0 \Big\}, \quad (2.27)\end{aligned}$$

where $\psi(s_1^-,...,s_N^-,s_1^+,...,s_M^+)$ is an appropriate aggregator of the individual additive factors, also often referred to as 'slacks'. Selecting different types of $\psi(s_1^-,...,s_N^-,s_1^+,...,s_M^+)$ will give different types of the slack-based measures or additive measures of technical efficiency, as those considered in detail by Tone (2001), Fukuyama and Weber (2009), Färe et al. (2015) and an application in Pham and Zelenyuk (2018), to mention a few.

2.3 The Rise of Allocative (In)efficiencies

The duality between the DDF and profit function, concisely in (2.9)-(2.10), implies that for any $p\in\Re^M_{++}$, $w\in\Re^N_{++}$ and for any pre-specified vector $d=(-d_x,d_y)\in\Re^N_-\times\Re^M_+$, we have

$$\mathcal{NPE}_d(x,y,p,w|\Psi)\geq\mathcal{DDF}_d(x,y|\Psi),\ (x,y)\in\Psi. \quad (2.28)$$

That is, there is always a non-negative (and possibly zero) gap between these two measures of efficiency. This gap, usually defined in this context as the *difference* between $\mathcal{NPE}_d$ and $\mathcal{DDF}_d$, is referred to as the directional allocative efficiency measure, which we will denote as $\mathcal{AE}_d(x, y, p, w|\Psi)$. It measures inefficiency due to a non-optimal allocation of inputs and outputs with respect to their prices, according to the profit-maximization criterion (1.2) and along the direction d.

In turn, this gives rise to the following decomposition, for all $(x, y) \in \Psi$, any prices $(w, p) \in \Re^N_{++} \times \Re^M_{++}$ and any direction $d = (-d_x, d_y) \in \Re^N_{-} \times \Re^M_{+}$, we have:

$$\mathcal{NPE}_d(x, y, p, w|\Psi) = \mathcal{DDF}_d(x, y|\Psi) + \mathcal{AE}_d(x, y, p, w|\Psi). \tag{2.29}$$

Furthermore, while the DDF has a duality relation with the profit efficiency, the output oriented Debreu-Farrell technical efficiency measure (or the output oriented Shephard's distance function) has a duality relation with the revenue efficiency.[8] Indeed, for any $p \in \Re^M_{++}$, we have

$$\mathcal{RE}(x, y, p|\Psi) \geq \mathcal{OTE}(x, y|\Psi), \ \forall (x, y) \in \Psi \tag{2.30}$$

and the gap between these two measures of efficiency, usually defined as the *ratio* between them, is referred to as the output oriented allocative efficiency measure, which we denote as $\mathcal{OAE}(x, y, p|\Psi)$, gives rise to the following de-

[8]Some of this duality theory requires $P(x)$ to be convex.

composition, for all $(x, y) \in \Psi$:

$$\mathcal{RE}(x, y, p|\Psi) = \mathcal{OTE}(x, y|\Psi) \times \mathcal{OAE}(x, y, p|\Psi). \tag{2.31}$$

Hence, $\mathcal{OAE}(x, y, p|\Psi)$ measures inefficiency due to a non-optimal allocation of outputs with respect to their prices, according to the revenue-maximization criterion (2.5).

On the other hand, the input oriented Debreu-Farrell technical efficiency measure is dual to the cost efficiency.[9] Specifically, for any $w \in \Re^N_{++}$, we have[10]

$$\mathcal{CE}(x, y, w|\Psi) \leq \mathcal{ITE}(x, y|\Psi), \ \forall (x, y) \in \Psi \tag{2.32}$$

and the gap between these two measures of efficiency, usually defined as the ratio between them, is referred to as the input oriented allocative efficiency measure, which we denote as $\mathcal{IAE}(x, y, w|\Psi)$. In turn, this gives rise to the following decomposition

$$\mathcal{CE}(x, y, w|\Psi) = \mathcal{ITE}(x, y|\Psi) \times \mathcal{IAE}(x, y, w|\Psi), \ \forall (x, y) \in \Psi. \tag{2.33}$$

Note that this version of the allocative inefficiency measure, $\mathcal{IAE}(x, y, w|\Psi)$, measures inefficiency due to a non-optimal allocation of inputs with respect to their prices, according to the cost-minimization criterion (2.3).

[9] Some of this duality theory requires $L(y)$ to be convex.

[10] In the literature, (2.28), (2.30) as well as (2.32) are sometimes called Mahler-type inequalities.

It is worth noting another interesting discovery of Färe et al. (2019)—the decomposition of the profit efficiency measure (1.23) into three distinct components or sources of profit inefficiency: (i) output oriented technical efficiency, (ii) output allocative efficiency (call it $\mathcal{OAE}(x^j, y^j, p|\Psi)$) and (iii) a new allocative efficiency that measures the gap between profit maximization and revenue maximization, which we will denote as $\mathcal{RAE}(x^j, y^j, p, w|\Psi)$). Or, formally, we have

$$\begin{aligned}&\mathcal{OPE}(x^j, y^j; w, p|\Psi)\\&\quad= \mathcal{OTE}(x^j, y^j|\Psi) \times \mathcal{OAE}(x^j, y^j, p|\Psi) \times \mathcal{RAE}(x^j, y^j, p, w|\Psi), \qquad (2.34)\end{aligned}$$

where

$$\mathcal{RAE}(x^j, y^j, p, w|\Psi) = \frac{\mathcal{OPE}(x^j, y^j; w, p|\Psi)}{\mathcal{RE}(x^j, y^j, p|\Psi)}. \qquad (2.35)$$

Similar developments can be done for the input orientation of the profit efficiency, thus connecting it to the Debreu-Farrell input oriented efficiency measure.

Finally, combining (2.31) and (2.33) with (2.1), we have the following decomposition:

$$\begin{aligned}\mathcal{OPE}(x^j, y^j; w, p|\Psi) \quad &= \quad \frac{\mathcal{PF}(p, w|\Psi)}{r^j} + \\ &+ \mathcal{ITE}(x^j, y^j|\Psi) \times \mathcal{IAE}(x^j, y^j, w|\Psi) \\ &\times \mathcal{OTE}(x^j, y^j|\Psi) \times \mathcal{OAE}(x^j, y^j, p|\Psi) \\ &\times \mathcal{CBR}^*(x^j, y^j, w, p|\Psi). \end{aligned} \tag{2.36}$$

2.4 Merits and Limitations

Each of the discussed measures have their merits and limitations that make them advantageous in some contexts and disadvantageous in others. There is no one-fit-for-all measure.

For example, an advantage of the Russell-type measures and of the slack-based measures of efficiency is that they typically reach 'more optimal' solutions than the directional efficiency measures. This is because the latter are restricted to follow a particular direction of measurement, which may or may not project onto a point on the frontier that is also efficient in a more general sense. In particular, they may or may not indicate Pareto-Koopmans efficiency, as defined in (1.21). For example, imagine a two-dimensional input space: an allocation on a 'flat' portion of the input isoquant will be deemed efficient for most of the directions, yet there will still be a slack in one of the inputs that can be reduced while keeping the other input, output and technology constant.

On the other hand, the disadvantage of the Russell-type and the slack-based measures is that they are harder to interpret than, for example, the Debreu-Farrell efficiency measures, besides also being harder to compute. Moreover, they may also have discontinuity issues, especially when estimated with nonparametric approaches.[11] Meanwhile, the directional measures of efficiency are fairly intuitive, especially the Debreu-Farrell (or Shephard) measures, which express efficiency in percentage terms with clear implications of what can be expanded or contracted.

Finally, the dual measures of efficiency (i.e., cost, revenue and profit efficiency) can be viewed as superior to the corresponding primal efficiency measures as they also account for the allocative aspect with respect to prices of inputs or outputs, if such information is available and reliable. They also take into account the slacks in the selected orientation (for cost and revenue efficiencies), or even all the possible slacks and all the allocative inefficiencies (for profit efficiency). On the other hand, the price information might be unavailable or unreliable, which is why researchers often resort to technical (or primal) efficiency measurements or a hybrid of primal and dual measures (e.g., see Zelenyuk (2020b)).

[11] E.g., see Russell (1990), Dmitruk and Koshevoy (1991) and Sickles and Zelenyuk (2019, Chapter 3).

Chapter 3

Dynamics

3.1 Index Numbers Theory of Productivity in a Nutshell

In a simple case when there is only one output and one input, a researcher can simply use the so-called single factor productivity index (SFP), defined as the ratio of 'output per input ratios' of two different periods or, equivalently, as the ratio of outputs of different periods over the ratio of inputs of those periods, i.e.,

$$\mathcal{SFP}_{st} \equiv \frac{y_t/x_t}{y_s/x_s} = \frac{y_t/y_s}{x_t/x_s}, \tag{3.1}$$

where the subscripts s and t indicate specific time periods (with convention $s < t$).

A very popular example of (3.1) is the Labor Productivity Index (LPI), measuring the change in output (e.g., GDP) per unit of labor. While being a useful metric for measuring productivity dynamics, its limitation is that it effectively ignores the other inputs that helped in producing the output, besides it requiring a univariate measure of output.

To overcome these limitations, various approaches have been proposed in the literature. One approach is to basically use the same formula as (3.1), but replace the ratio of univariate outputs y_t/y_s with an output quantity index and replace the ratio of univariate inputs x_t/x_s with an input quantity index.

An important question then is which output and input indexes would be best for those roles? This natural question is a subject in itself and is at the core of the index numbers theory in general, spanning several centuries of research. And, there is no simple answer to this question, except, perhaps this general conclusion from one of the patriarchs of this field, made 100 years ago:

> “What index numbers are "best?" Naturally much depends on the purpose in view.”
>
> – Fisher (1921, p.533).

As a result, various alternatives were proposed in the literature. Currently, it appears that the most popular approach is based on the so-called Malmquist quantity indexes. As the name suggests, the idea of such indexes goes back to Malmquist (1953) and parallels the idea of the ‘true cost of living index’ from Konüs (1924). It was then theoretically solidified in the seminal *Econometrica* paper by Caves et al. (1982) and elaborated on later in Diewert (1992), which also gave rise to the so-called Hicks-Moorsteen Productivity Index (HMPI), which was further developed by Bjurek (1996) and others. The latter two papers also proposed, focused on and advocated a totally new (at that time) productivity index, which they dubbed as the Malmquist Productivity Index (MPI), while others refer to it as the 'CCD-index' approach. Meanwhile, the idea of the HMPI was more thoroughly developed and advocated by Bjurek (1996) and more recently by Briec and Kerstens (2004, 2011), Epure et al. (2011) and Mayer and Zelenyuk (2014), among others, while Färe et al. (1996), Mizobuchi (2017a,b) and Färe and Zelenyuk (2021a) compared the two indexes theoretically and/or empirically. In this chapter we will also focus our attention on these two indexes.

3.2 Intertemporal Technology Characterizations

To model the dynamic aspects of production processes in general and to define HMPI and MPI in particular, the theoretical definitions must be generalized to the intertemporal context. To do so, consider three arbitrary periods s, t, τ (some of which may coincide), which we will use in subscripts or superscripts to denote the relevant time periods. In particular, the *technology set in period* τ is defined in general terms as

$$\Psi^\tau = \{(x, y) \ : \ \ x \text{ can produce } y \text{ at period } \tau\} \tag{3.2}$$

and the regularity conditions discussed in Chapter 1 are adapted accordingly. All other mathematical objects we defined above ($P(x)$, $L(y)$, $\mathcal{R}(x,p)$, $\mathcal{C}(y,w)$, etc.) can be adapted accordingly. To save space, we will only do so for the Shephard distance functions (and hence for technical efficiency) context.[1]

Let us denote $\mathcal{ODF}(x_s, y_t|\Psi^\tau)$ as the Shephard output distance function for the input level observed in the period s, with the technology available in the same period τ, and given the output level observed in the period t, i.e.,

$$\mathcal{ODF}(x_s, y_t|\Psi^\tau) = \inf_\gamma\{\gamma > 0 \ : \ (x_s, y_t/\gamma) \in \Psi^\tau\}. \tag{3.3}$$

Similarly, we denote with $\mathcal{IDF}(x_s, y_t|\Psi^\tau)$ the Shephard input distance function for the output level observed in the period t, with the technology available in the same period τ, and given the input level observed in the period s, i.e.,

$$\mathcal{IDF}(x_s, y_t|\Psi^\tau) = \sup_\lambda\{\lambda > 0 \ : \ (x_s/\lambda, y_t) \in \Psi^\tau\}. \tag{3.4}$$

The essence of the productivity measurement is therefore to develop a 'coherent' or 'well-justified' way of measuring not only the efficiency (i.e., a distance between (x, y) and a frontier of technology) in different periods, but also how it changes over time and, as importantly, how technology changes

[1] More general formulations involving indexes based on revenue efficiency and cost efficiency can be found in, for example, Zelenyuk (2006), Mayer and Zelenyuk (2014), Sickles and Zelenyuk (2019) and references therein.

over time. To measure the latter phenomenon in practice, early studies (e.g., Hicks (1932)) have set a tradition of assuming the so-called 'Hicks-neutral technology change'—the notion we will consider in some detail in the next section.

3.3 Hicks-Neutrality

3.3.1 Single-output Case

In a simple context with a single-output, where technology can be characterized via a production function (2.20), the notion of Hicks-neutral technology change can be defined as a situation where

$$y = \mathcal{F}_\tau(x|\Psi^\tau) \;\; = \;\; \mathcal{A}(\tau) \times \mathcal{F}(x), \quad \forall x \in \Re^N_+. \tag{3.5}$$

That is, a Hicks-neutral production function is (multiplicatively) separable into two parts: $\mathcal{F}(x)$, which is the part that describes the transformation of inputs into outputs independently from time and a multiplicative residual part, $\mathcal{A}(\tau)$, that describes how technology changes over time.

Clearly, (3.5) is a very simplified (and somewhat naive) view on the technological progress, especially these days when some inputs are getting more (or less) intensively used in production than before, i.e., implying a biased technology change. Nevertheless, (3.5) has been used in many economic models, including the seminal works of Solow (1956, 1957) and many recent works after that, and so it is important to consider it here, along with its generalizations to multi-output-multi-input contexts.

The assumption of Hicks neutrality indeed helps simplifying (somewhat naively and possibly misleadingly) the modeling of productivity dynamics. For example, the simple productivity measure (3.1) can be rewritten as

$$\begin{aligned} \mathcal{SFP}_{st} &= \frac{\mathcal{F}_t(x_t|\Psi^t)/x_{lt}}{\mathcal{F}_s(x_s|\Psi^s)/x_{ls}} \\ &= \frac{\mathcal{F}(x_t)/x_{lt}}{\mathcal{F}(x_s)/x_{ls}} \times \frac{\mathcal{A}(t)}{\mathcal{A}(s)}, \end{aligned} \tag{3.6}$$

where the second line gives the decomposition of the simple productivity index into the dynamics due to the contributions from changes in all inputs (normalized by one of the inputs denoted as x_{ls} and x_{lt}, e.g., labor) and, as a multiplicative residual, the change in technology $\mathcal{A}(t)/\mathcal{A}(s)$.

With the additional assumption of differentiability of $\mathcal{F}_\tau$, one can get further decomposition of the first component into contributions from changes in each input (weighted by their importance in the production), which was the idea of the seminal work of Solow (1957). Specifically, we can get

$$\mathfrak{g}(y_t) = \sum_{j=1}^{N} e_{jt}\mathfrak{g}(x_{jt}) + \mathfrak{g}(\mathcal{A}(t)), \tag{3.7}$$

where $\mathfrak{g}(z_t)$ denotes the growth rate of z_t (i.e., $\mathfrak{g}(z_t) := (dz_t/dt)/z_t = d\ln(z_t)/dt \approx (z_{t+\triangle t}/z_t) - 1$) and $e_{jt} \equiv (\partial\mathcal{F}(x_t)/\partial x_{jt}) \times (x_{jt}/\mathcal{F}(x_t))$ is the *partial scale elasticity* w.r.t. input j. Hence, we have a decomposition of growth in output into contribution from each input and from technology change (also referred to as 'Solow residual'). Moreover, if constant returns to scale is assumed then (due to linear homogeneity of $\mathcal{F}(x_t)$ under CRS), y_t and each x_{jt} can be normalized by one of the inputs, e.g., typically by labor (say x_{lt}), and so we get decomposition of the growth in labor productivity:

$$\mathfrak{g}(y_t/x_{lt}) \quad = \sum_{j=1, j\neq l}^{N} e_{jt}\mathfrak{g}(x_{jt}/x_{lt}) + \mathfrak{g}(\mathcal{A}(t)). \tag{3.8}$$

This approach became one of the main workhorses in applied work on productivity for many decades, under the brand of *growth accounting*. The more modern approaches of productivity measurement can be viewed as generalizations of this approach that attempt to resolve some of its caveats and limitations. For example, the standard growth accounting assumed that all observations (e.g., countries) are efficient, in the sense that the observed y_τ is the same as $\mathcal{F}_\tau(x|\Psi^\tau)$. However, if we admit (for whatever reason) a possibility of some inefficiency at a period τ, denoted by $e_\tau \in (0,1]$ so that $y_\tau = e_\tau \times \mathcal{F}_\tau(x|\Psi^\tau)$, then the decomposition of single factor productivity index (3.6) can be extended further as

$$\mathcal{SFP}_{st} \quad = \quad \frac{\mathcal{F}(x_t)/x_{lt}}{\mathcal{F}(x_s)/x_{ls}} \times \frac{\mathcal{A}(t)}{\mathcal{A}(s)} \times \frac{e_t}{e_s}, \tag{3.9}$$

where the third term is an index measuring the change in efficiency (e.g., measured via output oriented Shephard's distance function).

3.3.2 Multi-output Case

A generalization of the simple Hicks-neutrality (3.5) can be made in various ways, which are essentially the generalizations of the notion of homotheticity (or multiplicative separability) for the intertemporal characterizations.

Here we will follow the generalized definitions from Sickles and Zelenyuk (2019).[2] Specifically, technology change between some periods s and t is called *Hicks output-neutral* if and only if the *output set* in the period t can be multiplicatively decomposed into the output set in the period s and a scaling factor representing technology change, i.e.,

$$P^t(x) = \mathcal{A}_O(x, s, t)P^s(x),\ \forall x \in \Re^N_+. \tag{3.10}$$

Similarly, technology is called *Hicks input-neutral* if and only if the input set in the period t can be multiplicatively decomposed into the input set in the period s and a scaling factor representing a change in technology, i.e.,

$$L^t(y) = \mathcal{A}_I(y, s, t)L^s(y),\ \forall y \in \Re^M_+. \tag{3.11}$$

Note that these generalizations of (3.5) are local in the sense that they depend on s and t as well as allow for the output scaling factor to vary with x for (3.10), and for the input scaling factor to vary with y for (3.11). A more restrictive form of Hicks neutrality can also be (and often is) assumed, where

[2] For more details, see Blackorby et al. (1976), Chambers and Färe (1994) and most recently in Färe et al. (2020).

these scaling factors are independent from x and y, i.e., $\mathcal{A}_O(x, s, t) = \mathcal{A}_O(s, t)$ and $\mathcal{A}_I(y, s, t) = \mathcal{A}_I(s, t)$. Even more restrictive form is when $\mathcal{A}_O(s, t) = 1/\mathcal{A}_I(s, t)$.

Once assumed for the technology, they imply equivalent structures on its primal and dual characterizations. For example, (3.10) is equivalent to

$$\mathcal{ODF}(x, y|\Psi^t) = \frac{\mathcal{ODF}(x, y|\Psi^s)}{\mathcal{A}_O(x, s, t)}, \ \forall (x, y) \in \Re^N_+ \times \Re^M_+, \tag{3.12}$$

while (3.11) is equivalent to

$$\mathcal{IDF}(x, y|\Psi^t) = \frac{\mathcal{IDF}(x, y|\Psi^s)}{\mathcal{A}_I(y, s, t)}, \ \forall (x, y) \in \Re^N_+ \times \Re^M_+. \tag{3.13}$$

Again, while being fairly restrictive assumptions on technology, they are sometimes assumed to simplify the modeling and measurement of productivity dynamics, as we will discuss in the next sections.

3.4 Malmquist Quantity Indexes

The following four indexes are often used for measuring changes in aggregate quantities of outputs and inputs:

- Malmquist *output* quantity index w.r.t. period s (or Laspeyres perspective) as the reference is defined as

$$Q_y(y_s, y_t, x_s|\Psi^s) \equiv \frac{\mathcal{ODF}(x_s, y_t|\Psi^s)}{\mathcal{ODF}(x_s, y_s|\Psi^s)}. \tag{3.14}$$

- Malmquist *output* quantity index w.r.t. period t (or Paasche perspective) as the reference is defined as

$$Q_y(y_s, y_t, x_t|\Psi^t) \equiv \frac{\mathcal{ODF}(x_t, y_t|\Psi^t)}{\mathcal{ODF}(x_t, y_s|\Psi^t)}. \tag{3.15}$$

- Malmquist *input* quantity index w.r.t. period s (or Laspeyres perspective) as the reference is defined as

$$Q_x(x_s, x_t, y_s|\Psi^s) \equiv \frac{\mathcal{IDF}(x_t, y_s|\Psi^s)}{\mathcal{IDF}(x_s, y_s|\Psi^s)}. \tag{3.16}$$

- Malmquist *input* quantity index w.r.t. period t (or Paasche perspective) as the reference is defined as

$$Q_x(x_s, x_t, y_t|\Psi^t) \equiv \frac{\mathcal{IDF}(x_t, y_t|\Psi^t)}{\mathcal{IDF}(x_s, y_t|\Psi^t)}. \tag{3.17}$$

Because these indexes are designed to compare two periods, they are referred to as *bilateral* indexes. Importantly, note that the results from these indexes may depend on the choice of the reference, s or t. Indeed, (3.14) can potentially be very different from (3.15), while (3.16) can be potentially very different from (3.17). This is especially the case when there is substantial length between s and t. To see this more formally, note that the output quantity indexes (3.14)–(3.15) are independent from the reference inputs *if and only if* the technology is *output homothetic*, i.e., (1.14) holds, because this structure will be equivalent to

$$Q_y(y_s, y_t, x_s|\Psi^s) = \frac{\mathcal{ODF}(1_N, y_t|\Psi^s)}{\mathcal{ODF}(1_N, y_s|\Psi^s)}$$

and

$$Q_y(y_s, y_t, x_t|\Psi^t) = \frac{\mathcal{ODF}(1_N, y_t|\Psi^t)}{\mathcal{ODF}(1_N, y_s|\Psi^t)}.$$

Moreover, if the change in technology between s and t is Hicks-neutral in the output space, i.e., if (3.10) holds, then we would have

$$Q_y(y_s, y_t, x_s|\Psi^s) = Q_y(y_s, y_t, x_t|\Psi^t). \tag{3.18}$$

Analogously, the input quantity indexes are independent from the reference outputs *if and only if* the technology is *input homothetic*, i.e., (1.15) holds, and so

$$Q_x(x_s, x_t, y_s|\Psi^s) = \frac{\mathcal{IDF}(x_t, 1_M|\Psi^s)}{\mathcal{IDF}(x_s, 1_M|\Psi^s)}$$

and

$$Q_x(x_s, x_t, y_t|\Psi^t) = \frac{\mathcal{IDF}(x_t, 1_M|\Psi^t)}{\mathcal{IDF}(x_s, 1_M|\Psi^t)}.$$

Moreover, if technological change is Hicks-neutral in the input space, i.e., (3.11) holds, then we have

$$Q_x(x_s, x_t, y_s|\Psi^s) \quad = \quad Q_x(x_s, x_t, y_t|\Psi^t). \tag{3.19}$$

For more detailed discussions about homotheticity and Hicks neutrality and their implications to productivity measurements, see Blackorby et al. (1976); Färe and Mitchell (1993); Chambers and Färe (1994, 1998); Färe et al. (2020) and references therein.

3.5 Hicks-Moorsteen Productivity Indexes

In their turn, and while being useful for various reasons, the Malmquist quantity indexes serve as building blocks for the following productivity indexes:

Hicks-Moorsteen *productivity* index w.r.t. period s (or Laspeyres perspective)

$$\mathcal{HMPI}(x_s, x_t, y_s, y_t|\Psi^s) \equiv \frac{Q_y(y_s, y_t, x_s|\Psi^s)}{Q_x(x_s, x_t, y_s|\Psi^s)}, \tag{3.20}$$

Hicks-Moorsteen *productivity* index w.r.t. period t (or Paasche perspective)

$$\mathcal{HMPI}(x_s, x_t, y_s, y_t|\Psi^t) \equiv \frac{Q_y(y_s, y_t, x_t|\Psi^t)}{Q_x(x_s, x_t, y_t|\Psi^t)}. \tag{3.21}$$

Importantly, because the results from Malmquist quantity indexes may depend on the choice of the reference, s or t, so can the results from these HMPIs. That is, (3.20) can be potentially very different from (3.21). This is particularly the case when there is substantial length between s and t.

Indeed, using data from the seminal work of Kumar and Russell (2002) (originally from Penn World Tables), Färe and Zelenyuk (2021a) pointed out the dramatic differences between the conclusions from these two perspectives and discussed theoretical justifications for the aggregation of them.[3]

The simplest (and theoretically justified) way to avoid the arbitrariness, which researchers usually take to reconcile the Laspeyres and Paasche perspectives, is the Fisher's approach by taking the *simple geometric* mean of them. This gives the (averaged over the references) Hicks-Moorsteen *productivity* index, defined as

$$\begin{aligned}
&\mathcal{HMPI}(x_s, x_t, y_s, y_t|\Psi^s, \Psi^t) \\
&\equiv \left[\mathcal{HMPI}(x_s, x_t, y_s, y_t|\Psi^s) \times \mathcal{HMPI}(x_s, x_t, y_s, y_t|\Psi^t)\right]^{\frac{1}{2}} \\
&= \left[\frac{Q_y(y_s, y_t, x_s|\Psi^s)}{Q_x(x_s, x_t, y_s|\Psi^s)} \times \frac{Q_y(y_s, y_t, x_t|\Psi^t)}{Q_x(x_s, x_t, y_t|\Psi^t)}\right]^{\frac{1}{2}}. \qquad (3.22)
\end{aligned}$$

A natural question arises: Can one simplify this (or another) index by just fixing the reference, say at some Ψ^0, and relate all measurements to that common reference? For example, this could help with the so-called 'transitivity' (or 'circularity') of the index. Of course one can! The problem is whether that would give a meaningful measurement. Indeed, an implicit principle here is that the reference for the measurement should be 'relevant' for the objects that are measured. The exact meaning of what is 'relevant'

[3] Also see Färe and Zelenyuk (2019) for analogous discussions on 'additive indexes' or indicators, e.g. based on directional distance functions.

generally may depend on the context and hence is infeasible to specify for all the possibilities and, in fact, may also involve some subjective value judgements (which should be made transparent and motivated). For example, a motivation for why Ψ^s and Ψ^t (and hence their complete characterizations that enter (3.20), (3.21) and (3.22), or their dual analogues) are relevant can be based on the fact that those were the possibility sets faced by the economic agents at those times when they were making decisions. Meanwhile, choosing some other reference, Ψ^0, may be unfair if it characterized a substantially different set of possibilities, and thus potentially misleading the measurements and resulting policy implications.

More discussions on the virtues and limitations of these indexes and their relation to other indexes (including those based on directional distance functions) can be found in Briec and Kerstens (2004, 2011); Epure et al. (2011), Chapter 4 of Sickles and Zelenyuk (2019), Färe and Zelenyuk (2021a) and references therein.

3.6 Malmquist Productivity Indexes

The *Malmquist Productivity Index* (MPI), or the CCD-index approach, is another and apparently the most popular approach for measuring productivity dynamics for multi-output-multi-input frameworks.

3.6.1 Output and Input Orientations of MPI

For the output orientation, Caves et al. (1982) proposed using the output oriented period-s (or Laspeyres perspective) Malmquist Productivity Index,

$$\mathcal{OMPI}(x_s, x_t, y_s, y_t|\Psi^s) \equiv \frac{\mathcal{ODF}(x_t, y_t|\Psi^s)}{\mathcal{ODF}(x_s, y_s|\Psi^s)}, \tag{3.23}$$

and when choosing technology in the period t as the reference (or Paasche perspective), they suggested the output oriented period-t Malmquist Productivity Index,

$$\mathcal{OMPI}(x_s, x_t, y_s, y_t|\Psi^t) \equiv \frac{\mathcal{ODF}(x_t, y_t|\Psi^t)}{\mathcal{ODF}(x_s, y_s|\Psi^t)}. \tag{3.24}$$

Similarly as for the HMPI, to avoid dependency on the choice of the time reference, researchers usually follow Fisher's approach by taking the (simple) *geometric* mean of the two perspectives, giving the output oriented Malmquist Productivity Index (MPI):

$$\begin{aligned}
&\mathcal{OMPI}(x_s, x_t, y_s, y_t|\Psi^s, \Psi^t) \\
&\quad \equiv \left[\mathcal{OMPI}(x_s, x_t, y_s, y_t|\Psi^s) \times \mathcal{OMPI}(x_s, x_t, y_s, y_t|\Psi^t)\right]^{\frac{1}{2}} \\
&\quad = \left[\frac{\mathcal{ODF}(x_t, y_t|\Psi^s)}{\mathcal{ODF}(x_s, y_s|\Psi^s)} \times \frac{\mathcal{ODF}(x_t, y_t|\Psi^t)}{\mathcal{ODF}(x_s, y_s|\Psi^t)}\right]^{\frac{1}{2}}.
\end{aligned} \tag{3.25}$$

Furthermore, recall that the output oriented Shephard's distance function is just one of the possible ways to characterize technology. Basically, one could use any other alternatives to it and obtain an alternative index defined in a similar fashion. For example, when choosing the input oriented Shephard's distance function, then we get the input oriented s-period (or Laspeyres perspective) Malmquist Productivity Index,

$$\mathcal{IMPI}(y_s, y_t, x_s, x_t|\Psi^s) \equiv \frac{\mathcal{IDF}(y_t, x_t|\Psi^s)}{\mathcal{IDF}(y_s, x_s|\Psi^s)}, \tag{3.26}$$

and, the input oriented t-period (or Paasche perspective) Malmquist Productivity Indexes,

$$\mathcal{IMPI}(y_s, y_t, x_s, x_t|\Psi^t) \equiv \frac{\mathcal{IDF}(y_t, x_t|\Psi^t)}{\mathcal{IDF}(y_s, x_s|\Psi^t)}, \tag{3.27}$$

while the simple geometric mean (or Fisher's perspective) of these two input oriented MPIs is given by

$$\begin{aligned} &\mathcal{IMPI}(y_s, y_t, x_s, x_t|\Psi^s, \Psi^t) \\ &\quad \equiv \left[\mathcal{IMPI}(y_s, y_t, x_s, x_t|\Psi^s) \times \mathcal{IMPI}(y_s, y_t, x_s, x_t|\Psi^t)\right]^{\frac{1}{2}} \\ &\quad = \left[\frac{\mathcal{IDF}(y_t, x_t|\Psi^s)}{\mathcal{IDF}(y_s, x_s|\Psi^s)} \times \frac{\mathcal{IDF}(y_t, x_t|\Psi^t)}{\mathcal{IDF}(y_s, x_s|\Psi^t)}\right]^{\frac{1}{2}}. \end{aligned} \tag{3.28}$$

Importantly, note that under the assumption of CRS (which is very common for aggregate cross-country studies), we have an equivalence between the output oriented MPI and the input oriented MPI, in the sense that

$$\mathcal{OMPI}(x_s, x_t, y_s, y_t|\Psi^s, \Psi^t) = \frac{1}{\mathcal{IMPI}(y_s, y_t, x_s, x_t|\Psi^s, \Psi^t)} \tag{3.29}$$

for any (x_s, y_s) and (x_t, y_t) in their allowed domains. This equality follows directly from (2.22) (which is equivalent to the definition of CRS), when applying to the definitions of $\mathcal{OMPI}$ and $\mathcal{IMPI}$. It does not hold in general for any technology, although it may still hold approximately if technology is 'not far' from CRS.

3.6.2 Decomposition of MPI

A particularly appealing feature of this productivity index is the possibility of its decomposition into various sub-indexes. The most popular decomposition of MPI is into two sources: one measuring the *efficiency change*, while the other measuring the *technology change*. It was popularized by the seminal work of Färe et al. (1994) and accomplished as following

$$\begin{aligned}
&\mathcal{OMPI}(x_s, x_t, y_s, y_t|\Psi^s, \Psi^t) \\
&\quad = \mathcal{OEC}(x_s, x_t, y_s, y_t|\Psi^s, \Psi^t) \times \mathcal{OTC}(x_s, x_t, y_s, y_t|\Psi^s, \Psi^t),
\end{aligned} \tag{3.30}$$

where the first component in the r.h.s. of (3.30) is the index measuring the efficiency change, defined as

$$\mathcal{OEC}(x_s, x_t, y_s, y_t|\Psi^s, \Psi^t) \equiv \frac{\mathcal{ODF}(x_t, y_t|\Psi^t)}{\mathcal{ODF}(x_s, y_s|\Psi^s)}, \tag{3.31}$$

while the second component in the r.h.s. of (3.30) is the index measuring the (average) technology change, defined as

$$\mathcal{OTC}(x_s, x_t, y_s, y_t|\Psi^s, \Psi^t) \equiv \left[\frac{\mathcal{ODF}(x_t, y_t|\Psi^s)}{\mathcal{ODF}(x_t, y_t|\Psi^t)} \times \frac{\mathcal{ODF}(x_s, y_s|\Psi^s)}{\mathcal{ODF}(x_s, y_s|\Psi^t)}\right]^{\frac{1}{2}}. \tag{3.32}$$

Note that (3.30) is a generalization of (3.6) and (3.9) that not only allows for inefficiency (as also does (3.9)) but also does not impose the Hicks-neutrality of technology change, in addition to being general enough to embrace multi-output-multi-input context. In principle, it also does not require CRS, although such an assumption is often considered natural for measuring productivity dynamics, especially on a macroeconomic level. A notable work in this vein is that of Kumar and Russell (2002), who used a sample of countries from Penn World Tables, considering their aggregate single-output variable (real GDP in international dollars) and two inputs, labour (L) and capital (K), for 1965 vs. 1990. They estimated the labor productivity index (LPI) and decomposed it into three sources, as in (3.9), but without the assumption of Hicks-neutrality, and without assuming differentiability or any particular form of the production function. Due to the single-output-two-input framework, their decomposition was

$$\mathcal{SFP}_{st} \;=\; \mathcal{OEC}_{st} \times \mathcal{OTC}_{st} \times \mathcal{KLACC}_{st},$$

where $\mathcal{OEC}_{st}$ and $\mathcal{OTC}_{st}$ are as those in (3.31) and (3.32), while $\mathcal{KLACC}_{st}$ is a measure of accumulation of capital per labor, also referred to as the 'capital deepening' contribution to the productivity change. Kumar and Russell (2002) and the later papers they inspired found that it is this last component, $\mathcal{KLACC}_{st}$, that was responsible for most of the change in the labor productivity (and income per worker) dynamics. The so-called Asian Tigers were among the most prominent examples of this phenomenon, where the massive capital per worker increase led to sky-rocketing economic growth from 1960s to 1990s.[4]

3.7 Equivalence of Different Indexes

While in general the conceptually different productivity indexes, like MPI and HMPI, may potentially yield very different results and conclusions, they often give similar estimates and also may coincide exactly under certain conditions on technology. These conditions typically involve a combination of CRS, homotheticity, Hicks-neutrality of technology change assumptions. For example, combining the results from the previous sections ((1.14)-(1.15) and

[4]For other applications involving this and related decompositions, see Henderson and Russell (2005), Henderson and Zelenyuk (2007) and Badunenko et al. (2008) and a recent comprehensive review of this literature by Badunenko et al. (2017).

(3.10)-(3.13)), we have

$$\mathcal{OMPI}(x_s, x_t, y_s, y_t|\Psi^\tau) = \mathcal{HMPI}(x_s, x_t, y_s, y_t|\Psi^\tau) \tag{3.33}$$

$\forall(x_s, y_s),\ (x_t, y_t) \in \Re_+^N \times \Re_+^M$ if and only if technology in both periods ($\tau = s, t$) is inverse homothetic and exhibits CRS.[5]

Moreover, under other and more specific (parametric) assumptions on technology, they are also equivalent to the practical statistical indexes such as the Fisher productivity index (FPI), Törnqvist productivity index (TPI), etc. Specifically, if one is willing to assume the revenue (or profit) maximizing and cost minimizing behavior (in both s and t) and that the technology is characterized by the Shephard output distance functions that has the translog functional form with the second-order coefficients being equal for periods s and t, i.e., for $\tau = s, t$ along with some other assumptions, then the MPI is equivalent to the Törnqvist productivity index. Meanwhile, if instead of the translog, the 'Diewert functional form' is assumed (along with some other assumptions), then the MPI and HMPI are equivalent to the Fisher productivity index.[6]

Importantly, note that even if the MPI and HMPI as well as the FPI

[5]See Färe et al. (1996) for the proof. Also see Mizobuchi (2017a,b) and Färe et al. (2020) for more recent related discussions and further references.

[6]These results go back to the classic works of Caves et al. (1982) and Diewert (1992). Also see textbook discussions in Balk (1998, 2008) and Chapter 7 of Sickles and Zelenyuk (2019), who provided more references. Some further generalizations to quadratic mean of order-r indexes (including Walsh index, etc.) can be found in Mizobuchi and Zelenyuk (2020).

and TPI are yielding very similar results, they could be very different from the results suggested by the more widely used labor productivity index, potentially implying different policy implications (e.g., see Färe and Zelenyuk (2021a) for a real data example and discussions).

Finally, it is worth emphasizing that the MPI and HMPI, as well as all the measures of efficiency we considered in the previous chapter, are sometimes referred to as 'true' (rather than empirical) measures in the analogous sense as the 'true cost of living index' from Konüs (1924). That is, they are based on the abstract concepts or principles of economic theory and are unobserved by a researcher and need to be estimated in some way, e.g., via DEA, SFA or other methods. Badunenko et al. (2017) provides a comprehensive review of papers applying such indexes to analyze 'Productivity of Nations', while Sickles and Zelenyuk (2019) gives an extensive methodological discussion of virtues and caveats; both works also provide many key references on this topic.

Chapter 4

Aggregation

So far, an implicit focus has been on measuring the efficiency of an individual, such as a firm or more generally a *decision making unit* (DMU). Another important question is about the efficiency of a larger system that consists of individual DMUs, which one wants to consider explicitly as the parts composing the group. This is an important question and it is a sub-stream of the literature by itself, spanning economic theory, operations research, statistics and empirical work. In some sense, this question is analogous to other aggregation questions that have been raised many times in economics and, in fact, it is also at the core of index numbers literature.

In the context of efficiency analysis, this important question was addressed starting from, at least as early as, Farrell (1957) who introduced a new (at that time) concept of 'Structural Efficiency of an Industry', defined as the weighted average of the individual efficiency scores. In particular, Farrell (1957) proposed aggregating his input oriented efficiency scores by weighting them with the output shares of the corresponding DMUs, i.e.,

$$\overline{\mathcal{ITE}} = \sum_{k=1}^{n} \mathcal{ITE}(x^k, y^k|\Psi)S^k, \tag{4.1}$$

where

$$S^k = y^k / \sum_{i=1}^{n} y^i. \tag{4.2}$$

While the idea was very intuitive, some issues or unexplained questions remained. For example: Why exactly were the output shares chosen as weights to aggregate the efficiency scores, especially since they were input oriented efficiency scores? And, as importantly: What should one do in the case of several outputs? These questions, as well as the general aggregation question, were further researched by many researchers from different angles in the past four decades, e.g., by Førsund and Hjalmarsson (1979), Li and Ng (1995), Blackorby and Russell (1999), Färe and Zelenyuk (2003), Briec et al. (2003), Li and Cheng (2007), Färe et al. (2004), Färe and Zelenyuk (2005), Bogetoft and Wang (2005), Zelenyuk (2006), Mussard and Peypoch (2006), Färe and Zelenyuk (2007), Cooper et al. (2007a), Simar and Zelenyuk (2007), Nesterenko and Zelenyuk (2007), Färe et al. (2008), Pachkova (2009), Kuosmanen et al. (2010), Raa (2011), Mayer and Zelenyuk (2014, 2019), Färe and Karagiannis (2014), Karagiannis (2015), Karagiannis and Lovell (2015), to mention a few. While it is impossible to cover all these works here in detail, it is worth outlining some of the key results, which are at the core of many of these works.[1]

4.1 Koopmans Aggregation Theorem

The key result used for developing different theories of aggregation in productivity and efficiency analysis goes back to Koopmans (1957), and we will refer to it as the Koopmans aggregation theorem. This theorem starts with an assumption that the aggregate technology set of a group of say n firms can be defined as the summation of the individual technology sets of all n DMUs in this group, i.e.,[2]

$$\Psi^* \equiv \sum_{\oplus k=1}^{n} \Psi^k. \tag{4.3}$$

Note that this aggregation structure assumes that there are no (net) externalities in the aggregation across the individual technology sets. If externalities were present then the aggregate technology may be different from Ψ^*, larger or smaller, depending on the type (positive or negative), and on the extent and particular features of the externalities. For example, in reality there might be some dissonance effect among some DMUs, while also

[1] This chapter is also a brief and updated synthesis of more detailed discussions found in Sickles and Zelenyuk (2019, Chapter 5), Mayer and Zelenyuk (2019) and Zelenyuk (2020a), which were cast at different angles, where the reader is referred to for more information and related references.

[2] To be precise, it is the Minkowski summation, which is a generalization of the usual summation, developed for summing sets as collections of vectors. Formally, for any sets A and B, the Minkowski summation of A and B, sometimes conventionally denoted as $A \oplus B$, is defined as a new set that contains all the possible summations of any of the element of A with any of the elements of B, i.e., $A \oplus B := \{a + b : a \in A, b \in B\}$ (e.g., see Krein and Smulian (1940) and Oks and Sharir (2006)). In economics, the Minkowski summation was also used in the Shapley–Folkman-Starr theorem (Starr (2008)) and related results.

synergy effects among others, and any of them may or may not dominate the other at a particular time or circumstance, which could also change over time. Predicting a particular structure of externalities and how it changes over time is hardly feasible in general terms and is well-beyond the scope of this manuscript. Hence, we adopt (4.3) or its analogs that we discuss below, as has been used in the literature so far.

Once the aggregate technology is defined, all the objects that we have defined above, such as output and input sets, profit, revenue and cost functions, distance functions, etc., can then be defined with respect to this aggregate technology, analogously how they were defined above for the individual technologies. For example, the *aggregate profit function* of the group of $k = 1, ..., n$ DMUs can be defined analogously to (1.2), yet on the aggregate technology (4.3), i.e.,

$$\mathcal{PF}(p, w|\Psi^*) = \sup_{x,y}\{py - wx \ : \ (x, y) \in \Psi^*\}. \tag{4.4}$$

The Koopmans aggregation theorem then establishes an equilibrium result stating that under the same prices (p, w), the aggregate profit (4.4), when the group optimizes with respect to (4.3), coincides with the summation of the individual profits (1.2) where each individual optimized independently (and assuming no externalities) with respect to its technology Ψ^k, i.e.,

$$\overline{\mathcal{PF}}(p, w|\Psi^*) = \sum_{k=1}^{n} \mathcal{PF}(p, w|\Psi^k). \tag{4.5}$$

4.2 Aggregation of Profit Efficiency and its Decomposition

The aggregation theorem about the profit function (4.5) can be used to derive the aggregation result for various profit efficiencies. For example, if we define the *Aggregate Nerlovian (or directional) profit efficiency measure* as

$$\overline{\mathcal{NPE}}_d(\sum_{j=1}^{n} x^j, \sum_{j=1}^{n} y^j, p, w|\Psi^*) \equiv \frac{\overline{\mathcal{PF}}(p, w|\Psi^*) - (p\sum_{j=1}^{n} y^j - w\sum_{j=1}^{n} x^j)}{(pd_y + wd_x)} \tag{4.6}$$

for any $d = (-d_x, d_y)$ which is a non-zero vector in $\Re^N_- \times \Re^M_+$ such that $pd_y + wd_x \neq 0$, then we have the aggregation result for the Nerlovian profit efficiencies:

$$\overline{\mathcal{NPE}}_d(\sum_{j=1}^{n} x^j, \sum_{j=1}^{n} y^j, p, w|\Psi^*) = \sum_{k=1}^{n} \mathcal{NPE}_d(x^k, y^k, p, w|\Psi^k). \tag{4.7}$$

Furthermore, from the duality theory applied to the aggregate technology, we get the Mahler-type inequality for the aggregate Nerlovian profit efficiency, i.e., for all $(x, y) \in \Psi^*$, all $(w, p) \in \Re^{N+M}_{++}$ and $(-d_x, d_y) \in \Re^N_- \times \Re^M_+ \backslash \{0_{N+M}\}$, we have

$$\overline{\mathcal{NPE}}_d(\sum_{j=1}^{n} x^j, \sum_{j=1}^{n} y^j, p, w|\Psi^*) \geq \overline{\mathcal{DDF}}_d(\sum_{j=1}^{n} x^j, \sum_{j=1}^{n} y^j|\Psi^*), \tag{4.8}$$

where the r.h.s. of (4.8) is the usual directional distance function but defined w.r.t. the aggregate technology Ψ^* and evaluated at $(\sum_{j=1}^n x^j, \sum_{j=1}^n y^j)$, i.e.,

$$\overline{\mathcal{DDF}}_d(\sum_{j=1}^n x^j, \sum_{j=1}^n y^j|\Psi^*) \equiv \sup_{\beta}\{\beta \in \Re \ : \ (\sum_{j=1}^n x^j, \sum_{j=1}^n y^j) + \beta d \in \Psi^*\}. \quad (4.9)$$

Similarly as in the disaggregate case, one can close the inequality gap in (4.8) by introducing an additive residual representing a *directional aggregate allocative efficiency*, which we denote as $\overline{\mathcal{AE}}_d(\sum_{j=1}^n x^j, \sum_{j=1}^n y^j, p, w|\Psi^*)$, thus leading to the following decomposition

$$\begin{aligned}\overline{\mathcal{NPE}}_d(\sum_{j=1}^n x^j, \sum_{j=1}^n y^j, p, w|\Psi^*) = {} & \overline{\mathcal{DDF}}_d(\sum_{j=1}^n x^j, \sum_{j=1}^n y^j|\Psi^*) \\ & + \overline{\mathcal{AE}}_d(\sum_{j=1}^n x^j, \sum_{j=1}^n y^j, p, w|\Psi^*), \qquad (4.10)\end{aligned}$$

which holds for all $(x, y) \in \Psi^*$, all $(w, p) \in \Re^{N+M}_{++}$ and any $(-d_x, d_y) \in \Re^N_- \times \Re^M_+ \backslash \{0_{N+M}\}$ such that $pd_y + wd_x \neq 0$.

Hence, combining (4.7) with (2.29), we arrive at

$$\begin{aligned}\overline{\mathcal{NPE}}_d(\sum_{j=1}^n x^j, \sum_{j=1}^n y^j, p, w|\Psi^*) = {} & \sum_{j=1}^n \mathcal{DDF}_d(x^j, y^j|\Psi^j) \\ & + \sum_{j=1}^n \mathcal{AE}_d(x^j, y^j, p, w|\Psi^j), \qquad (4.11)\end{aligned}$$

and therefore, we have

$$\overline{\mathcal{DDF}}_d(\sum_{j=1}^n x^j, \sum_{j=1}^n y^j|\Psi^*) = \sum_{j=1}^n \mathcal{DDF}_d(x^j, y^j|\Psi^j) \qquad (4.12)$$

if and only if

$$\overline{\mathcal{AE}}_d(\sum_{j=1}^{n} x^j, \sum_{j=1}^{n} y^j, p, w|\Psi^*) = \sum_{j=1}^{n} \mathcal{AE}_d(x^j, y^j, p, w|\Psi^j). \tag{4.13}$$

Leveraging on the theory of Pexider equations, Färe et al. (2008) also pointed out that such conditions imply that the DDF must possess an affine functional form, while also pointing out that, more generally, we have

$$\overline{\mathcal{DDF}}_d(\sum_{j=1}^{n} x^j, \sum_{j=1}^{n} y^j|\Psi^*) \geq \sum_{j=1}^{n} \mathcal{DDF}_d(x^j, y^j|\Psi^j) \tag{4.14}$$

and

$$\overline{\mathcal{AE}}_d(\sum_{j=1}^{n} x^j, \sum_{j=1}^{n} y^j, p, w|\Psi^*) \leq \sum_{j=1}^{n} \mathcal{AE}_d(x^j, y^j, p, w|\Psi^j). \tag{4.15}$$

That is, for the same directional vector, the sum of individual DDFs provides a lower bound for the aggregate DDF, while the sum of individual directional allocative efficiencies provides an upper bound for the directional aggregate allocative efficiencies.

4.3 Aggregation for the Output Orientation

For the output oriented case, and following Färe and Zelenyuk (2003), we have a revenue analogue of the Koopmans aggregation theorem. Specifically, we have an equilibrium relationship between the aggregate maximal revenue and the individual maximal revenues, i.e.,

$$\overline{\mathcal{R}}(x^1, ..., x^n, p|\overline{P}(x^1, ..., x^n)) = \sum_{k=1}^{n} \mathcal{R}(x^k, p|\Psi^k), \tag{4.16}$$

for any $p \in \mathbb{R}^M_{++}$, where on the l.h.s. we have the aggregate *revenue function,* defined as

$$\overline{\mathcal{R}}(x^1, ..., x^n, p|\overline{P}(x^1, ..., x^n)) \;\; = \;\; \sup_y \{py \;:\; y \in \overline{P}(x^1, ..., x^n)\}, \tag{4.17}$$

which, note, is optimized over $\overline{P}(x^1, ..., x^n)$ rather than Ψ^*, which is the *aggregate output set* defined as the Minkowski summation of the individual output sets, i.e.,[3]

$$\overline{P}(x^1, ..., x^n) = \sum_{\oplus k=1}^{n} P^k(x^k|\Psi^k). \tag{4.18}$$

Moreover, the associated aggregate revenue efficiency, an analogue to (2.4), is then given by

$$\begin{aligned} &\overline{\mathcal{RE}}(x^1, ..., x^n, \sum_{k=1}^{n} y^k, p|\overline{P}(x^1, ..., x^n)) \\ &\quad = \frac{\overline{\mathcal{R}}(x^1, ..., x^n, p|\overline{P}(x^1, ..., x^n))}{p\sum_{k=1}^{n} y^k}. \end{aligned} \tag{4.19}$$

The equilibrium relation (4.16) is then used to obtain the aggregation result for relating the aggregate revenue efficiency (4.19) to the individual

[3]Note that this assumes there are no (net) externalities associated with the aggregation across the individual output sets.

revenue efficiencies (2.4), given by

$$\overline{\mathcal{RE}}(x^1,...,x^n,\sum_{k=1}^{n}y^k,p|\overline{P}(x^1,...,x^n)) = \sum_{k=1}^{n}\mathcal{RE}(x^k,y^k,p|\Psi^k)S^k, \qquad (4.20)$$

where for the weight we have S^k, which is the revenue share of the individual DMU k in the group, i.e.,

$$S^k = \frac{py^k}{p\sum_{k=1}^{n}y^k}. \qquad (4.21)$$

Furthermore, with the help of some algebra, the r.h.s. of the aggregation result (4.20) is then decomposed into the aggregates of technical and allocative efficiencies, i.e.,

$$\overline{\mathcal{RE}}(x^1,...,x^n,\sum_{k=1}^{n}y^k,p|\overline{P}(x^1,...,x^n)) = \overline{\mathcal{OTE}} \times \overline{\mathcal{OAE}}, \qquad (4.22)$$

where

$$\overline{\mathcal{OTE}} \equiv \sum_{k=1}^{n}\mathcal{OTE}(x^k,y^k|\Psi^k) \times S^k, \qquad (4.23)$$

and

$$\overline{\mathcal{OAE}} \equiv \sum_{k=1}^{n}\mathcal{OAE}(x^k,y^k,p|\Psi^k) \times S^k_{ae}, \qquad (4.24)$$

where the weights are given by

$$S^k_{ae} \equiv \frac{p(y^k\mathcal{OTE}(x^k,y^k|\Psi^k))}{p\sum_{k=1}^{n}(y^k\mathcal{OTE}(x^k,y^k|\Psi^k))},\ \ k=1,...,n. \qquad (4.25)$$

It is important to note that this aggregation scheme is not the only possible way to obtain an aggregate efficiency from the individual efficiency scores and other alternatives are also possible. One of such alternatives is the sample mean, which is a standard descriptive statistic that remains useful and it is important to present it too, contrasting to one where some justified economic weights are factored in the averaging. An advantage of the sample mean is its simplicity–all the weights are $1/n$, where n is the sample size. An advantage of the the aggregate efficiencies presented above is that they account for an economic weight of each DMU k in the sample, doing so in an intuitive way, with weights derived from some economic principles.

In practice, looking at both non-weighted and weighted means can be very important whenever the group of interest consists of DMUs of very different sizes, which is very common in reality. Indeed, many industries are often dominated by the 'big four' or 'top 5' and so on, and ignoring such substantial inequality in their size or 'mass' across DMUs is, in a sense, analogous to ignoring the mass of chemicals (e.g., molar mass) when analyzing substances composed of chemicals of different mass.

Importantly, since the sample mean of individual efficiency scores basically ignores an economic weight or size of the individuals being aggregated, the aggregate efficiency and the sample mean can yield estimates and conclusions that are very different, potentially leading to opposite policy implications. For example, think of an industry dominated by a few very inefficient (e.g., only 50% efficient) 'too-big-to-fail' entities that control (in all the respects, including political influence) say 80% of all the industry, operating along with hundreds of small firms that compete fiercely among themselves for their survival, attaining about 100% efficiency, yet altogether 'weigh' only 20% of the industry. The sample mean of efficiency scores for such an industry would suggest it is fairly efficient, while the aggregate efficiency with the weights 80/20 will suggest that this industry is very inefficient, or possibly requiring certain reforms, perhaps urgently needed to mitigate a crisis rooted in such high inequality combined with high inefficiency.[4]

4.4 Aggregation for Productivity Indexes

The Koopmans-type aggregation theory can also be extended to the dynamic context. To do so, we resort to the concept of *aggregate revenue-based Malmquist productivity index* or aggregate CCD-index (Zelenyuk (2006)), defined for measuring changes in productivity between s and t, as

$$\begin{aligned}
&\overline{\mathcal{RM}}(x_s^1, ..., x_s^n, x_t^1, ..., x_t^n, \sum_{k=1}^{n} y_s^k, \sum_{k=1}^{n} y_t^k, p_s, p_t) \\
&\equiv \left[\frac{\overline{\mathcal{RE}}(x_t^1, ..., x_t^n, \sum_{k=1}^{n} y_t^k, p_t | \overline{P}_s(x_s^1, ..., x_s^n))}{\overline{\mathcal{RE}}(x_s^1, ..., x_s^n, \sum_{k=1}^{n} y_s^k, p_s | \overline{P}_s(x_s^1, ..., x_s^n))} \times \right. \\
&\quad \left. \times \frac{\overline{\mathcal{RE}}(x_t^1, ..., x_t^n, \sum_{k=1}^{n} y_t^k, p_t | \overline{P}_t(x_t^1, ..., x_t^n))}{\overline{\mathcal{RE}}(x_s^1, ..., x_s^n, \sum_{k=1}^{n} y_s^k, p_s | \overline{P}_t(x_t^1, ..., x_t^n))}\right]^{-\frac{1}{2}}, \qquad (4.26)
\end{aligned}$$

[4]See Färe and Zelenyuk (2003) and Simar and Zelenyuk (2007, 2018, 2020) for related discussions and detailed derivations.

where the time subscripts indicate the particular values of efficiency measures for specific periods s, t.

Then, a solution to the aggregation problem is obtained by applying the *dynamic version* of the revenue analogue to the Koopmans aggregation theorem (Zelenyuk (2006)), where for all $x^k \in \Re^N_+,\ \forall k = 1, ..., n,\ p \in \Re^M_{++}, \tau = s, t$, we have

$$\overline{\mathcal{R}}(x^1, ..., x^n, p|\overline{P}_\tau(x^1, ..., x^n)) = \sum\nolimits_{k=1}^{n} \mathcal{R}^k(x^k, p|\Psi^k_\tau), \tag{4.27}$$

and therefore

$$\overline{\mathcal{RE}}(x^1_j, ..., x^n_j, \sum_{k=1}^{n} y^k_j, p_j|\overline{P}_\tau(x^1, ..., x^n)) = \sum\nolimits_{k=1}^{n} \mathcal{RE}^k(x^k_j, y^k_j, p_j|\Psi^k_\tau) \times S^k_j, \tag{4.28}$$

where S^k_j is a dynamic version of (4.21), for $j = s, t$, defined as

$$S^k_j \equiv \frac{p_j y^k_j}{p_j \sum_{k=1}^{n} y^k_j},\ k = 1, ..., n. \tag{4.29}$$

Moreover, we also get the decomposition into the technical and allocative aggregates

$$\overline{\mathcal{RE}}(x^1_j, ..., x^n_j, \sum_{k=1}^{n} y^k_j, p_j|\overline{P}_\tau(x^1, ..., x^n)) = \overline{\mathcal{OTE}}(j|\tau) \quad \times \quad \overline{\mathcal{OAE}}(j|\tau), \tag{4.30}$$

where

$$\overline{\mathcal{OTE}}(j|\tau) \equiv \sum\nolimits_{k=1}^{n} \mathcal{OTE}^k(x^k_j, y^k_j|\Psi^k_\tau) \times S^k_j, \tag{4.31}$$

and

$$\overline{\mathcal{OAE}}(j|\tau) \equiv \sum\nolimits_{k=1}^{n} \mathcal{OAE}^k(x_j^k, y_j^k, p_j|\Psi_\tau^k) \times S_{ae,\tau,j}^k, \tag{4.32}$$

where

$$S_{ae,\tau,j}^k \equiv \frac{p_j\left(y_j^k \mathcal{OTE}^k(x_j^k, y_j^k|\Psi_\tau^k)\right)}{p_j \sum_{k=1}^{n}\left(y_j^k \mathcal{OTE}^k(x_j^k, y_j^k|\Psi_\tau^k)\right)},\ k = 1, ..., n. \tag{4.33}$$

Applying (4.28) to each of the four components of (4.26), leads to

$$\begin{aligned}
&\overline{\mathcal{RM}}(x_s^1, ..., x_s^n, x_t^1, ..., x_t^n, \sum_{k=1}^{n} y_s^k, \sum_{k=1}^{n} y_t^k, p_s, p_t) \\
&\quad = \Bigg[\frac{\sum_{k=1}^{n} \mathcal{RE}^k(x_t^k, y_t^k, p_t|\Psi_s^k) \times S_t^k}{\sum_{k=1}^{n} \mathcal{RE}^k(x_s^k, y_s^k, p_s|\Psi_s^k) \times S_s^k} \\
&\quad\quad \times \frac{\sum_{k=1}^{n} \mathcal{RE}^k(x_t^k, y_t^k, p_t|\Psi_t^k) \times S_t^k}{\sum_{k=1}^{n} \mathcal{RE}^k(x_s^k, y_s^k, p_s|\Psi_t^k) \times S_s^k}\Bigg]^{-1/2}.
\end{aligned} \tag{4.34}$$

Furthermore, engaging the decomposition (4.30) to each component of (4.34), we get the analogous decomposition of the aggregate revenue-based MPI into the aggregate technical MPI and aggregate allocative MPI, namely

$$\overline{\mathcal{RM}}(x_s^1, ..., x_s^n, x_t^1, ..., x_t^n, \sum_{k=1}^{n} y_s^k, \sum_{k=1}^{n} y_t^k, p_s, p_t) = \overline{\mathcal{OMPI}} \times \overline{\mathcal{OAMPI}}$$

where

$$\overline{\mathcal{OMPI}} = \left[\frac{\overline{\mathcal{OTE}}(t|s)}{\overline{\mathcal{OTE}}(s|s)} \times \frac{\overline{\mathcal{OTE}}(t|t)}{\overline{\mathcal{OTE}}(s|t)}\right]^{-1/2}, \tag{4.35}$$

and

$$\overline{\mathcal{OAMPI}} = \left[\frac{\overline{\mathcal{OAE}}(t|s)}{\overline{\mathcal{OAE}}(s|s)} \times \frac{\overline{\mathcal{OAE}}(t|t)}{\overline{\mathcal{OAE}}(s|t)}\right]^{-1/2}, \tag{4.36}$$

where, in turn, the four components inside (4.35) are given in (4.31) while the four components inside (4.36) are given in (4.32). A similar aggregation strategy can be done for a decomposition of MPI into efficiency change and technology change (Zelenyuk (2006)).

A similar intuition as that given at the end of the previous section applies here, with the remark that the weights now reflect the changes of the 'mass' of the DMUs over time, doing so from both the Laspeyres and Paasche perspectives, i.e., using both s-period and t-period weights.

4.5 Other Aggregation Results

The results briefly outlined above are just a few of what are currently known in this sub-area of research. In particular, and first of all, note that besides the aggregation for the directional distance function and the Nerlovian profit efficiency, we only considered the output orientation. Similar results hold for the input orientation, e.g., see Färe et al. (2004) for the efficiencies and Mayer and Zelenyuk (2019) for the Hicks-Moorsteen productivity indexes. Furthermore, a link to the geometric aggregation can also be established (e.g., see Färe and Zelenyuk (2005) and Zelenyuk (2006)), while the price-independent weights can also be derived with additional assumptions (e.g., Färe and Zelenyuk (2007)). Similar frameworks were also developed for the aggregation of scale elasticities (Färe and Zelenyuk (2012)) and for the aggregation of scale efficiencies (Zelenyuk (2015)).

Moreover, a generalization to the context of more than one group, where the aggregation can be done first within each group and then further aggregated over several groups into a larger group, was developed by Simar and Zelenyuk (2007). The latter work also made the first attempt to develop a statistical foundation for weighted averages of efficiency scores, for their estimation, bias correction and for statistical inference based on the subsampling bootstrap approach (for the context of DEA estimation). More recently, Simar and Zelenyuk (2018) derived the new central limit theorems (CLTs) for such estimated aggregate efficiency scores. Meanwhile, Simar and Zelenyuk (2020) proposed further improvements for the finite-sample approximation of these and other new CLTs via improving the estimation of the variance. Recent applications of these approaches to analyze the efficiency of hospitals include Nguyen and Zelenyuk (2021a,b).

A separate stream of literature on aggregation is dedicated to the notion of *reallocation* across DMUs of a group. Indeed, it is important to note that, unlike (4.3), the aggregation structure (4.18) implies that the aggregate output set presumes the re-allocation of inputs among the DMUs is *not* allowed. As a result, it is a subset of the output set implied by the more general aggregation structure (4.3), i.e., for any $(x^1, ..., x^n)$ we have

$$\overline{P}(x^1, ..., x^n) \subseteq P^*(\sum_{k=1}^{n} x^k), \tag{4.37}$$

where $P^*(\sum_{k=1}^{n} x^k)$ is the output set corresponding to the aggregate technology Ψ^*, i.e.,

$$P^*(\sum_{k=1}^{n} x^k) \equiv \{y : (\sum_{k=1}^{n} x^k, y) \in \Psi^*\}. \tag{4.38}$$

This was pointed out (and proven) by Nesterenko and Zelenyuk (2007), who then used it to introduce the notion of aggregate reallocative efficiency, designed to measure the gap between $\overline{P}(x^1, ..., x^n)$ and $P^*(\sum_{k=1}^{n} x^k)$, and decomposed it into technical and allocative aggregates, similar to how it has been described in previous sections (with a bit more complexity added). These ideas parallel the related and alternative developments in Bogetoft and Wang (2005), Pachkova (2009), Kuosmanen et al. (2010) and Raa (2011), and are further extended to the aggregation of Malmquist Productivity Indexes by Mayer and Zelenyuk (2014) and Mayer and Zelenyuk (2019), to mention a few. Similar developments can also be done for the cases when there are externalities that would define an aggregate technology set that is different from $P^*(\sum_{k=1}^{n} x^k)$ and from $\overline{P}(x^1, ..., x^n)$ and the resulting gaps can be measured in a similar fashion, thus facilitating a measurement of the inefficiency due to externalities.

Concluding Remarks

The goal of this monograph was to give a brief yet rigorous overview of the economic theory foundations for the field of Efficiency and Productivity Analysis. While it has not covered everything (which is really infeasible), I hope it provides enough of a primary foundation for some researchers to prepare and proceed more comfortably when reading past and current literature in this area, whether for a theoretical or for an empirical work, and involving any estimator—DEA, SFA or a symbiosis of these two or others.

In particular, if the reader is interested in a much deeper understanding of the theory, then reading the original papers cited here (and therein) would be the best option. A narrower focus would provide a deeper understanding however, it could also prove to be quite tedious and so this work will hopefully be a good primer for those journeys. If instead the reader is interested in a somewhat deeper, yet also broader coverage, then a natural step would be to learn from Färe and Primont (1995) or/and Chapters 1 through 7 of Sickles and Zelenyuk (2019), which also provide more references.

If the reader is interested in applications using SFA, then, given the myriad of estimators there, going through a primer that broadly reviews the area of SFA would perhaps be the best start. Among the recent works in such category are Kumbhakar et al. (2020a,b) and Chapter 11 through 16 of Sickles and Zelenyuk (2019). Meanwhile, a more time-constrained reader may find Sickles et al. (2020) more feasible time-wise (also covering bits of DEA), especially if they are interested in implementations via R or Matlab.[5]

If the reader is interested in applications using DEA, then, again given the myriad of different methods within DEA, starting with a primer might be more convenient and this is the goal of Part 2 of the project that yielded this manuscript. In the meantime, a recent concise review of DEA can be found in Ray (2020) while more extensive recent treatment can be found in Chapter 8 through 10 of Sickles and Zelenyuk (2019).[6]

Whichever direction is chosen, I hope this work will help you (a dear reader) to make interesting discoveries, whether in theory or in practice, within this wonderful and important field of research , as well as for its intersections with other fields.

[5] Also see classic works of Aigner et al. (1977), Schmidt and Sickles (1984), Greene (2005a,b, 2007) to mention a few, and the special issue in Journal of Econometrics edited by Kumbhakar and Schmidt (2016) for many interesting discussions.

[6] Also see classic works of Farrell (1957), Charnes et al. (1978), Boussofiane et al. (1991), Charnes et al. (1994), Førsund and Sarafoglou (2002), Cooper et al. (2007b), Emrouznejad et al. (2008), Cook and Seiford (2009) and Cooper et al. (2011b), Simar and Wilson (2013) to mention a few.

Acknowledgements

The author acknowledges the support from the University of Queensland and from the Australian Research Council (ARC Future Fellowship grant FT170100401) and thanks the Editors, referees and all others who gave feedback to various versions of this manuscript, including Anne-Claire Bouton, Bao Hoang Nguyen, Ruby Nguyen, Evelyn Smart, Zhichao Wang. These individuals and organizations are not responsible for the views expressed here.

Subject Index

Bibliography

Aigner, D., Lovell, C., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6(1):21–37.

Aparicio, J., Pastor, J. T., and Ray, S. C. (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.

Badunenko, O., Henderson, D. J., and Zelenyuk, V. (2008). Technological change and transition: Relative contributions to worldwide growth during the 1990s. *Oxford Bulletin of Economics and Statistics*, 70(4):461–492.

Badunenko, O., Henderson, D. J., and Zelenyuk, V. (2017). The productivity of nations. In Grifell-Tatjé, E., Lovell, C. A. K., and Sickles, R. C., editors, *The Oxford Handbook of Productivity*, chapter 8. New York, NY: Oxford University Press.

Balk, B. M. (1998). *Industrial Price, Quantity, and Productivity Indices:*

The Micro-Economic Theory and an Application. Boston, MA: Kluwer Academic Publishers.

Balk, B. M. (2008). *Price and Quantity Index Numbers: Models for Measuring Aggregate Change and Difference.* New York, NY: Cambridge University Press.

Bjurek, H. (1996). The Malmquist total factor productivity index. *The Scandinavian Journal of Economics*, 98(2):303–313.

Blackorby, C., Lovell, C. A. K., and Thursby, M. C. (1976). Extended Hicks neutral technical change. *The Economic Journal*, 86(344):845–852.

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

Bogetoft, P. and Wang, D. (2005). Estimating the potential gains from mergers. *Journal of Productivity Analysis*, 23(2):145–171.

Boussofiane, A., Dyson, R. G., and Thanassoulis, E. (1991). Applied data envelopment analysis. *European Journal of Operational Research*, 52(1):1–15.

Briec, W., Dervaux, B., and Leleu, H. (2003). Aggregation of directional distance functions and industrial efficiency. *Journal of Economics*, 79(3):237–261.

Briec, W. and Kerstens, K. (2004). 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. and Kerstens, K. (2011). The Hicks-Moorsteen productivity index satisfies the determinateness axiom. *The Manchester School*, 79(4):765–775.

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

Chambers, R., Chung, Y., and Färe, R. (1998). Profit, directional distance functions, and Nerlovian efficiency. *Journal of Optimization Theory and Applications*, 98(2):351–364.

Chambers, R. G., Chung, Y., and Färe, R. (1996). Benefit and distance functions. *Journal of Economic Theory*, 70(2):407–419.

Chambers, R. G. and Färe, R. (1994). Hicks neutrality and trade biased growth: A taxonomy. *Journal of Economic Theory*, 64(2):554–567.

Chambers, R. G. and Färe, R. (1998). Translation homotheticity. *Economic Theory*, 11(3):629–641.

Charnes, A., Cooper, W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6):429–444.

Charnes, A., Cooper, W. W., Lewin, A. Y., and Seiford, L. M., editors (1994). *Data Envelopment Analysis: Theory, Methodology, and Applications*. Amsterdam, NL: Springer.

Cook, W. and Seiford, L. (2009). Data envelopment analysis (DEA) - thirty years on. *European Journal of Operational Research*, 192(1):1–17.

Cooper, W., Huang, Z., Li, S., Parker, B., and Pastor, J. (2007a). Efficiency aggregation with enhanced Russell measures in data envelopment analysis. *Socio-Economic Planning Sciences*, 41(1):1–21.

Cooper, W., Pastor, J. T., Aparicio, J., and Borras, F. (2011a). Decomposing profit inefficiency in DEA through the weighted additive model. *European Journal of Operational Research*, 212(2):411 – 416.

Cooper, W. W., Seiford, L. M., and Tone, K. (2007b). *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*. New York, NY: Springer.

Cooper, W. W., Seiford, L. M., and Zhu, J., editors (2011b). *Handbook on Data Envelopment Analysis*. New York, NY: Springer.

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(3):481–507.

Diewert, W. E. (1973). Functional forms for profit and transformation functions. *Journal of Economic Theory*, 6(3):284–316.

Diewert, W. E. (1974). Applications of duality theory. In Intriligator, M. D. and Kendrick, D. A., editors, *Frontiers in Quantitative Economics*, volume 2, pages 106–171. Amsterdam, NL: Elsevier.

Diewert, W. E. (1982). Duality approaches to microeconomic theory. In *Handbook of Mathematical Economics*, volume 2, pages 535–599. Netherlands, NL: Elsevier.

Diewert, W. E. (1992). Fisher ideal output, input, and productivity indexes revisited. *Journal of Productivity Analysis*, 3(3):211–248.

Dmitruk, A. V. and Koshevoy, G. A. (1991). On the existence of a technical efficiency criterion. *Journal of Economic Theory*, 55(1):121–144.

Emrouznejad, A., Parker, B., and Tavares, G. (2008). Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in DEA. *Journal of Socio-Economics Planning Science*, 42(3):151–157.

Epure, M., Kerstens, K., and Prior, D. (2011). Technology-based total factor productivity and benchmarking: New proposals and an application. *Omega*, 39(6):608–619.

Färe, R., Førsund, F. R., Grosskopf, S., Hayes, K., and Heshmati, A. (2001).

A note on decomposing the Malmquist productivity index by means of subvector homotheticity. *Economic Theory*, 17(1):239–245.

Färe, R., Fukuyama, H., Grosskopf, S., and Zelenyuk, V. (2015). Decomposing profit efficiency using a slack-based directional distance function. *European Journal of Operational Research*, 247(1):335–337.

Färe, R. and Grosskopf, S. (1983). Measuring congestion in production. *Zeitschrift für Nationalökonomie / Journal of Economics*, 43(3):257–271.

Färe, R. and Grosskopf, S. (2009). A comment on weak disposability in nonparametric production analysis. *American Journal of Agricultural Economics*, 91(2):535–538.

Färe, R., Grosskopf, S., Norris, M., and Zhang, Z. (1994). Productivity growth, technical progress, and efficiency change in industrialized countries. *American Economic Review*, 84(1):66–83.

Färe, R., Grosskopf, S., and Roos, P. (1996). On two definitions of productivity. *Economics Letters*, 53(3):269–274.

Färe, R., Grosskopf, S., and Zelenyuk, V. (2004). Aggregation of cost efficiency: Indicators and indexes across firms. *Academia Economic Papers*, 32(3):395–411.

Färe, R., Grosskopf, S., and Zelenyuk, V. (2007). Finding common ground: Efficiency indices. In Färe, R., Grosskopf, S., and Primont, D., editors, *Aggregation, Efficiency and Measurement*, pages 83–95. Boston, MA: Springer.

Färe, R., Grosskopf, S., and Zelenyuk, V. (2008). Aggregation of Nerlovian profit indicator. *Applied Economics Letters*, 15(11):845–847.

Färe, R., He, X., Li, S. K., and Zelenyuk, V. (2019). A unifying framework for Farrell profit efficiency measurement. *Operations Research*, 67(1):183-197.

Färe, R. and Karagiannis, G. (2014). A postscript on aggregate Farrell efficiencies. *European Journal of Operational Research*, 233(3):784–786.

Färe, R. and Lovell, C. K. (1978). Measuring the technical efficiency of production. *Journal of Economic Theory*, 19(1):150–162.

Färe, R. and Mitchell, T. (1993). Multiple outputs and "homotheticity". *Southern Economic Journal*, 60(2):287–296.

Färe, R., Mizobuchi, H., and Zelenyuk, V. (2020). Hicks neutrality and homotheticity in technologies with multiple inputs and multiple outputs. *Omega*, forthcoming.

Färe, R. and Primont, D. (1995). *Multi-Output Production and Duality: Theory and Applications*. New York, NY: Kluwer Academic Publishers.

Färe, R. and Svensson, L. (1980). Congestion of production factors. *Econometrica: Journal of the Econometric Society*, 48(7):1745–1753.

Färe, R. and Zelenyuk, V. (2003). On aggregate Farrell efficiencies. *European Journal of Operational Research*, 146(3):615–620.

Färe, R. and Zelenyuk, V. (2005). On Farrell's decomposition and aggregation. *International Journal of Business and Economics*, 4(2):167–171.

Färe, R. and Zelenyuk, V. (2007). Extending Färe and Zelenyuk (2003). *European Journal of Operational Research*, 179(2):594–595.

Färe, R. and Zelenyuk, V. (2012). Aggregation of scale elasticities across firms. *Applied Economics Letters*, 19(16):1593–1597.

Färe, R. and Zelenyuk, V. (2019). On Luenberger input, output and productivity indicators. *Economics Letters*, 179:72 – 74.

Färe, R. and Zelenyuk, V. (2020). Profit efficiency: Generalization, business accounting and the role of convexity. *Economics Letters*, 196(C).

Färe, R. and Zelenyuk, V. (2021a). On aggregation of multi-factor productivity indexes. *Journal of Productivity Analysis*, 55(2):107–133.

Färe, R. and Zelenyuk, V. (2021b). Sequential data envelopment analysis. *Annals of Operations Research*, forthcoming.

Farrell, M. J. (1957). The measurement of productive efficiency. *Jornal of the Royal Statistical Society. Series A (General)*, 120(3):253–290.

Fisher, I. (1921). The best form of index number. *Journal of the American Statistical Association*, 17(133):533–537.

Førsund, F. R. and Hjalmarsson, L. (1979). Generalised Farrell measures of efficiency: An application to milk processing in Swedish dairy plants. *The Economic Journal*, 89(354):294–315.

Førsund, F. R. and Sarafoglou, N. (2002). On the origins of data envelopment analysis. *Journal of Productivity Analysis*, 17(1):23–40.

Foster, L., Haltiwanger, J., and Syverson, C. (2008). Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *American Economic Review*, 98(1):394–425.

Fukuyama, H. and Weber, W. (2009). A directional slacks-based measure of technical inefficiency. *Socio-Economic Planning Sciences*, 43(4):274–287.

Georgescu-Roegen, N. (1951). The aggregate linear production function and its applications to von Neumann's economic model. In Koopmans, T., editor, *Activity Analysis of Production and Allocation*. New York, NY: Wiley.

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. (2007). The econometric approach to efficiency analysis. In Fried, H. O., Lovell, C. A. K., and Schmidt, S. S., editors, *The Measure-*

ment of Productive Efficiency: Techniques and Applications. New York, NY: Oxford University Press.

Hanoch, G. (1970). Homotheticity in joint production. *Journal of Economic Theory*, 2(4):423–426.

Henderson, D. J. and Russell, R. R. (2005). Human capital and convergence: A production-frontier approach. *International Economic Review*, 46(4):1167–1205.

Henderson, D. J. and Zelenyuk, V. (2007). Testing for (efficiency) catching-up. *Southern Economic Journal*, 73(4):1003–1019.

Hicks, J. R. (1932). *The Theory of Wages*. London, UK: MacMillan.

Karagiannis, G. (2015). On structural and average technical efficiency. *Journal of Productivity Analysis*, 43(3):259–267.

Karagiannis, G. and Lovell, C. A. K. (2015). Productivity measurement in radial DEA models with a single constant input. *European Journal of Operational Research*, 251(1):323–328.

Konüs, A. A. (1939 [1924]). The problem of the true index of the cost of living. *Econometrica*, 7(1):10–29. Translated into English and published in 1939.

Koopmans, T. (1957). *Three Essays on The State of Economic Science*. New York, NY: McGraw-Hill.

Krein, M. and Smulian, V. (1940). On regulary convex sets in the space conjugate to a Banach space. *Annals of Mathematics*, 41(2):556–583.

Kumar, S. and Russell, R. R. (2002). Technological change, technological catch-up, and capital deepening: Relative contributions to growth and convergence. *American Economic Review*, 92(3):527–548.

Kumbhakar, S. and Schmidt, P. (2016). *Endogeneity Problems in Econometrics, Special Issue of the Journal of Econometrics*, volume 190. Amsterdam, Netherlands: Elsevier.

Kumbhakar, S. C. (1987). The specification of technical and allocative inefficiency in stochastic production and profit frontiers. *Journal of Econometrics*, 34(3):335–348.

Kumbhakar, S. C., Parmeter, C. F., and Zelenyuk, V. (2020a). *Stochastic Frontier Analysis: Foundations and Advances I*. Springer Singapore, Singapore. forthcoming.

Kumbhakar, S. C., Parmeter, C. F., and Zelenyuk, V. (2020b). *Stochastic Frontier Analysis: Foundations and Advances II*. Springer Singapore, Singapore. forthcoming.

Kuosmanen, T., Kortelainen, M., Sipiläinen, T., and Cherchye, L. (2010). Firm and industry level profit efficiency analysis using absolute and uniform shadow prices. *European Journal of Operational Research*, 202(2):584–594.

Lau, L. J. (1976). A characterization of the normalized restricted profit function. *Journal of Economic Theory*, 12(1):131–163.

Leibenstein, H. (1966). Allocative efficiency vs. "X-efficiency". *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. (1987). *Inside the Firm: The Inefficiencies of Hierarchy*. Cambridge, MA: Harvard University Press.

Leibenstein, H. and Maital, S. (1992). Empirical estimation and partitioning of X-inefficiency: A data-envelopment approach. *American Economic Review*, 82(2):428–433.

Li, S.-k. and Cheng, Y.-s. (2007). Solving the puzzles of structural efficiency. *European Journal of Operational Research*, 180(2):713–722.

Li, S.-K. and Ng, Y. C. (1995). Measuring the productive efficiency of a group of firms. *International Advances in Economic Research*, 1(4):377–390.

Malmquist, S. (1953). Index numbers and indifference surfaces. *Trabajos de Estadistica*, 4(2):209–242.

Mas-Colell, A., Whinston, M. D., and Green, J. R. (1995). *Microeconomic Theory*. Oxford University Press.

Mayer, A. and Zelenyuk, V. (2014). Aggregation of Malmquist productivity indexes allowing for reallocation of resources. *European Journal of Operational Research*, 238(3):774–785.

Mayer, A. and Zelenyuk, V. (2019). Aggregation of individual efficiency measures and productivity indices. In ten Raa, T. and Greene, W. H., editors, *The Palgrave Handbook of Economic Performance Analysis*, pages 527–557. Springer International Publishing, Cham.

McFadden, D. (1978). Cost, revenue, and profit functions. In McFadden, D. and Fuss, M. A., editors, *Production Economics: A Dual Approach to Theory and Applications*, volume 1: The Theory of Production, chapter 1. Amsterdam, NL: North-Holland.

Mizobuchi, H. (2017a). Productivity indexes under Hicks neutral technical change. *Journal of Productivity Analysis*, 48(1):63–68.

Mizobuchi, H. (2017b). A superlative index number formula for the Hicks-Moorsteen productivity index. *Journal of Productivity Analysis*, 48(2-3):167–178.

Mizobuchi, H. and Zelenyuk, V. (2020). Quadratic-mean-of-order-r indexes of output, input and productivity. CEPA Working Paper No. WP03/2020.

Mussard, S. and Peypoch, N. (2006). On multi-decomposition of the aggregate Malmquist productivity index. *Economics Letters*, 91(3):436–443.

Nesterenko, V. and Zelenyuk, V. (2007). Measuring potential gains from reallocation of resources. *Journal of Productivity Analysis*, 28(1–2):107–116.

Nguyen, B. H. and Zelenyuk, V. (2021a). Aggregate efficiency of industry and its groups: The case of Queensland public hospitals. *Empirical Economics*, forthcoming.

Nguyen, B. H. and Zelenyuk, V. (2021b). Aggregation of outputs and inputs for DEA analysis of hospital efficiency: Economics, operations research and data science perspectives. In Zhu, J. and Charles, V., editors, *Data-Enabled Analytics: DEA for Big Data*. Springer. forthcoming.

Oks, E. and Sharir, M. (2006). Minkowski sums of monotone and general simple polygons. *Discrete and Computational Geometry*, 35(2):223–240.

Pachkova, E. V. (2009). Restricted reallocation of resources. *European Journal of Operational Research*, 196(3):1049–1057.

Pham, M. D. and Zelenyuk, V. (2018). Slack-based directional distance function in the presence of bad outputs: theory and application to Vietnamese banking. *Empirical Economics*, 54(1):153–187.

Pham, M. D. and Zelenyuk, V. (2019). Weak disposability in nonparametric production analysis: A new taxonomy of reference technology sets. *European Journal of Operational Research*, 274(1):186 – 198.

Podinovski, V. V. and Kuosmanen, T. (2011). Modelling weak disposability in data envelopment analysis under relaxed convexity assumptions. *European Journal of Operational Research*, 211(3):577–585.

Raa, T. T. (2011). Benchmarking and industry performance. *Journal of Productivity Analysis*, 36(3):285–292.

Ray, S. C. (2020). *Data Envelopment Analysis: A Nonparametric Method of Production Analysis*, pages 1–62. Springer Singapore, Singapore.

Rubinstein, A. (2006). *Lecture Notes in Microeconomic Theory*. Number gradmicro1 in Online economics textbooks. SUNY-Oswego, Department of Economics.

Russell, R. R. (1990). Continuity of measures of technical efficiency. *Journal of Economic Theory*, 51(2):255–267.

Schmidt, P. and Sickles, R. C. (1984). Production frontiers and panel data. *Journal of Business & Economic Statistics*, 2(4):367–374.

Shephard, R. W. (1953). *Cost and Production Functions*. Princeton, NJ: Princeton University Press.

Shephard, R. W. (1970). *Theory of Cost and Production Functions*. Princeton studies in mathematical economics. Princeton, NJ: Princeton University Press.

Shephard, R. W. (1974). Indirect production functions. In *Mathematical Systems in Economics*, volume 10. Hain, Meisenheim am Glan.

Sickles, R. and Zelenyuk, V. (2019). *Measurement of Productivity and Efficiency: Theory and Practice*. New York, NY: Cambridge University Press.

Sickles, R. C., Song, W., and Zelenyuk, V. (2020). Chapter 8 - econometric analysis of productivity: Theory and implementation in R. In Vinod, H. D. and Rao, C., editors, *Financial, Macro and Micro Econometrics Using R*, volume 42 of *Handbook of Statistics*, pages 267 – 297. Elsevier.

Simar, L. and Wilson, P. W. (2013). Estimation and inference in nonparametric frontier models: Recent developments and perspectives. *Foundations and Trends in Econometrics*, 5(3–4):183–337.

Simar, L. and Zelenyuk, V. (2007). Statistical inference for aggregates of Farrell-type efficiencies. *Journal of Applied Econometrics*, 22(7):1367–1394.

Simar, L. and Zelenyuk, V. (2018). Central limit theorems for aggregate efficiency. *Operations Research*, 166(1):139–149.

Simar, L. and Zelenyuk, V. (2020). Improving finite sample approximation by central limit theorems for estimates from Data Envelopment Analysis. *European Journal of Operational Research*, 284(3):1002–1015.

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. *Review of Economics and Statistics*, 39(3):312–320.

Starr, R. M. (2008). Shapley-Folkman theorem. In Durlauf, S. N. and Blume, L. E., editors, *The New Palgrave Dictionary of Economics*, pages 317–318. Basingstoke, UK: Palgrave Macmillan.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving Decisions About Health, Wealth, and Happiness.* London, UK: Penguin.

Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research*, 130(3):498–509.

von Neumann, J. (1945). A model of general equilibrium. *Review of Economic Studies*, 13(1):1–9.

Zelenyuk, V. (2006). Aggregation of Malmquist productivity indexes. *European Journal of Operational Research*, 174(2):1076–1086.

Zelenyuk, V. (2013). A scale elasticity measure for directional distance function and its dual: Theory and DEA estimation. *European Journal of Operational Research*, 228(3):592–600.

Zelenyuk, V. (2014a). Directional vs. Shephard's distance functions. *Optimization Letters*, 8(8):2185–2189.

Zelenyuk, V. (2014b). Scale efficiency and homotheticity: Equivalence of primal and dual measures. *Journal of Productivity Analysis*, 42(1):15–24.

Zelenyuk, V. (2015). Aggregation of scale efficiency. *European Journal of Operational Research*, 240(1):269–277.

Zelenyuk, V. (2020a). Aggregation of efficiency and productivity: From firm to sector and higher levels. In Ray, S. C., Chambers, R., and Kumbhakar, S., editors, *Handbook of Production Economics*, pages 1–40. Springer Singapore, Singapore.

Zelenyuk, V. (2020b). Aggregation of inputs and outputs prior to data envelopment analysis under big data. *European Journal of Operational Research*, 282(1):172 – 187.