Contributions to Economics

Bert M. Balk

# Productivity

## Concepts, Measurement, Aggregation, and Decomposition

Springer

# Contributions to Economics

The series *Contributions to Economics* provides an outlet for innovative research in all areas of economics. Books published in the series are primarily monographs and multiple author works that present new research results on a clearly defined topic, but contributed volumes and conference proceedings are also considered. All books are published in print and ebook and disseminated and promoted globally. The series and the volumes published in it are indexed by Scopus and ISI (selected volumes).

More information about this series at http://www.springer.com/series/1262

Bert M. Balk

# Productivity

## Concepts, Measurement, Aggregation, and Decomposition

Springer

Bert M. Balk
Rotterdam School of Management
Erasmus University
Rotterdam, The Netherlands

ISSN 1431-1933 ISSN 2197-7178 (electronic)
Contributions to Economics
ISBN 978-3-030-75447-1 ISBN 978-3-030-75448-8 (eBook)
https://doi.org/10.1007/978-3-030-75448-8



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

# Preface

The main purpose of this book is to develop the theory of (intertemporal) productivity measurement without relying on the usual neo-classical assumptions, such as the existence of a production function characterized by constant returns to scale, optimizing behaviour of the economic agents, and perfect foresight. The theory can be applied to all the usual levels of aggregation (micro, meso, and macro), and half of the book is devoted to accounting for the links existing between the various levels. Basic insights from National Accounts are thereby used.

The book emphasizes the fundamental equivalence of multiplicative and additive models. The main instruments used are (multiplicative) indices and (additive) indicators. Throughout the book, the fundamental equivalence of these measurement tools is made explicit.

The final chapter is devoted to the decomposition of productivity change into the contributions of efficiency change, technological change, scale effects, and input- or output-mix effects. An application on a real-life data set shows the empirical feasibility of the theory.

The book is directed to a variety of overlapping audiences: statisticians involved in measuring productivity change (e.g., in national or international statistical agencies); economists interested in growth accounting; researchers relating macroeconomic productivity change to its industrial sources; micro-data researchers; and business analysts interested in performance measurement (through time or space).

The book tries to build a bridge between theory and practice. On the one hand, the textbook-orientated theorist is guided through a toolbox full of instruments, each with specific limitations and instructions for application. On the other hand, the practice-orientated policy developer is shown that the question "What is the (current) percentage of productivity change?" is easy to pose, but not so easy to answer. The book then provides the necessary background information.

The book can be seen as a sort of exposition of the programme sketched in "Chapter 2: Empirical productivity indices and indicators," in *The Oxford Handbook of Productivity Analysis*, edited by E. Grifell-Tatjé, C. A. K. Lovell, and R. C. Sickles, Oxford University Press, 2018. As the level of mathematics and economic theory never surpasses that of basic undergraduate courses, the book can be used for more advanced courses on measurement in economics, in academic settings, or elsewhere.

## Provenance of the Various Chapters

Nine of the ten chapters are based on previous publications. Most of these have been revised thoroughly, and consolidated to avoid overlap. The mathematical notation has been harmonized. References have been brought up-to-date, and extensions—some of which based on newer insights—have found an appropriate place. Nevertheless, readers familiar with the subject will recognize the sources of the various chapters. In principle, the chapters can be read independently. Here follows an overview.

- Section 1.1 is based on the corresponding section of "The residual: On monitoring and benchmarking firms, industries, and economies with respect to productivity", *Journal of Productivity Analysis* 20 (2003), 5–47.
- Chapter 2 is based on "An assumption-free framework for measuring productivity change", *The Review of Income and Wealth* 56 (2010), Special Issue 1, S224–S256, reprinted in *National Accounting and Economic Growth*, edited by John M. Hartwick, The International Library of Critical Writings in Economics No. 313 (Edward Elgar, Cheltenham UK, Northampton MA, 2016). Enhanced with parts of "Empirical Productivity Indices and Indicators," written in 2016, available at SSRN: http://ssrn.com/abstract=2776956. In particular Sect. 2.2.5, on the equivalence of multiplicative and additive models, is new. Added is a second Sect. 2.3.2 on growth accounting, and a Sect. 2.6 containing an empirical illustration of the approach developed in this chapter. Next, to make the book as self-contained as possible, Appendices A (on indices and indicators) and B (on decompositions of the value-added ratio) have been expanded. Appendix C, on the famous but frequently misunderstood Domar factor, is new.
- Chapter 3 is based on "Measuring and decomposing capital input cost", *The Review of Income and Wealth* 57 (2011), 490–512. Sections 3.3 (on the classic formulas) and 3.5 (on rates of return) have been expanded. Appendices A (Decompositions of time-series depreciation) and B (Geometric profiles) are new.
- Chapter 4 updates "Consistency issues in the construction of annual and quarterly productivity measures", *International Productivity Monitor* 37 (Fall 2019), 144–155.
- Chapter 5 expands "The dynamics of productivity change: A review of the bottom-up approach", in *Productivity and Efficiency Analysis*, edited by W. H. Greene, L. Khalaf, R. C. Sickles, M. Veall, and M.-C. Voia, Springer Proceedings in Business and Economics, Springer International Publishing Switzerland, 2016. In particular, Sect. 5.4.4, discussing cases where not all the data were accessible, is new. Section 5.5.2 (on the TRAD, CSLS, and GEAD decompositions) has been expanded. Sections 5.6 (on the logmean approach) and 5.7 (on the geometric and

harmonic approach) are new. Section 5.8 (on the monotonicity paradox) has been expanded. Finally, Appendices A (Reinsdorf's expansion of the GR method), B (Exercises on the Netherlands' manufacturing industry, 1984–1999), and C (Generalization of the OP decomposition) are new.

- Chapter 6 combines "Dissecting aggregate output and labour productivity change", *Journal of Productivity Analysis* 42 (2014), 35–43, and "Dissecting aggregate output and labour productivity change: A postscript on the role of relative prices" (with Jesus C. Dumagan), *Journal of Productivity Analysis* 45 (2015), 117–119. Appendix A (The Tang and Wang method) has been expanded.
- Chapter 7 is based on "Measuring and relating aggregate and subaggregate productivity change without neoclassical assumptions", *Statistica Neerlandica* 69 (2015), 21–48. Section 7.10, on growth accounting, is new.
- Chapter 8 is based on "A novel decomposition of aggregate total factor productivity change", *Journal of Productivity Analysis* 53 (2020), 95–105. Appendix A, analysing Baumol's "growth disease", is new.
- Chapter 9 updates "Aggregate productivity and productivity of the aggregate: Connecting the bottom-up and top-down approaches", in *Productivity and Inequality*, edited by W. H. Greene, L. Khalaf, P. Makdissi, R. C. Sickles, M. Veall, and M.-C. Voia, Springer Proceedings in Business and Economics, Springer International Publishing AG, 2018.
- Chapter 10 is entirely new, based on "The many decompositions of total factor productivity change" (with J. L. Zofío), Report No. ERS-2018-003-LIS, *Series Research in Management*, Erasmus Research Institute of Management, Erasmus University, Rotterdam, available at SSRN: http://ssrn.com/abstract=3167686.

As almost all the chapters are based on contributions to peer-reviewed journals and book collections, I want to thank all those colleagues who, in an entirely disinterested way, have spent time on reviewing initial versions of these chapters. The same applies to papers that did not make it to journals. I thank them for questions, criticism, and suggestions.

In particular, I thank José Luis Zofío for the most pleasant collaboration in the research underlying Chap. 10 and Paul de Boer for going through the almost-final version of the manuscript.

This book is dedicated to the Rotterdam School of Management, Erasmus University, especially its Department of Technology & Operations Management that continued to provide an intellectually stimulating environment after my formal retirement in 2011. Receive this book as a sort of farewell gift!

Amersfoort, The Netherlands — Bert M. Balk

# Contents

# List of Figures

# List of Tables

# Chapter 1
# Introduction

> Productivity … is a measure of the efficiency with which resources are converted into the commodities and services that men want.
>
> Solomon Fabricant, "Introduction", in Kendrick (1961).

> … productivity measurement is all about comparing outputs with inputs, …
>
> Ross Gittins, "Productivity should be a spin-free zone", The Sydney Morning Herald, June 23–24, 2007.

> Gains in productivity are the primary driver of wages and living standards.
>
> Sylvia Nasar, 2011, Grand Pursuit: The Story of Economic Genius (Fourth Estate, HarperCollinsPublishers, London)

> The traditional measure of the pace of innovation and technological change is total factor productivity (TFP) — output divided by a weighted average of labor and capital input.
>
> Robert J. Gordon (2016, 537)

## 1.1 The Origins of a Concept

As several distinguished scholars have written about the history of productivity measurement and analysis this introduction can be kept brief. The reader needs only sufficient information to place the contents of this book in a proper frame. For details the reader is referred to Griliches (2001), the first section of which is a reworked version of Griliches (1996) on "the discovery of the residual", Hulten (2001), Grifell-Tatjé and Lovell (2015, Chapter 1), and Grifell-Tatjé et al. (2018).

The first mention of total factor productivity (TFP) change as the ratio of an output quantity index and an input quantity index occurs in a contribution by Copeland (1937) to what, with hindsight, could be called the national income accounting approach. Stimulated by institutions such as the National Bureau of Economic Research, in the post-war period several studies were published, a typical one being Stigler (1947). These studies were mainly dealing with industry- or economy-wide aggregates. Although the TFP index was sometimes referred to as



B. M. Balk, *Productivity*, Contributions to Economics,
https://doi.org/10.1007/978-3-030-75448-8_1

a measure of the efficiency of the economic process, the common opinion was best voiced by Abramowitz (1956), who called it a "measure of our ignorance."[1]

The other, production-theoretic approach appears to go back to Tinbergen (1942). He extended the Cobb-Douglas production function with a time trend variable. The difference between the growth rate of real output and a weighted average of the growth rates of real capital and labour input was interpreted variably as efficiency change, technical development, or "Rationalisierungsgeschwindigkeit".

The basic and very influential contribution of Solow (1957) can be conceived as a sort of linkage of both traditions. He showed that under certain conditions the parameters of the Cobb-Douglas production function could be equated to observable statistical magnitudes, and the residual interpreted in terms of a ratio of output and input quantity index numbers. This is why the TFP index came to be known as the "Solow residual", although the name "residual" appears to have been used by Domar (1961) for the first time. Solow interpreted the residual as a measure of technical change.[2]

Since the inception of the concept of TFP change there have been two main lines of research. The first was directed at explanation. The second was directed at better measurement, primarily of the input factors capital and labour. In the beginning, the second style was more prominent than the first. For example, Jorgenson and Griliches (1967) claimed that using the "correct" index number framework and the "right" measurement of inputs would largely eliminate the role of the residual.

The residual disappeared indeed, but not at all due to better measurement techniques. The economy-wide disappearance of productivity growth in the seventies, its reappearance later on, and the search for the factors behind this world-wide phenomenon came to be known as the "productivity slowdown discussion". The emphasis shifted from measurement problems to explanation, and Griliches' work provides a clear demonstration of this shift. The main explanatory factors he considered were education, R&D expenditures, and patents.

The measurement problems, however, remained important. Looking back at a life-long of research in this area, Griliches (2001) said:

> It is my hunch that at least part of what happened [namely, the productivity slowdown] is that the economy and its various technological thrusts moved into sectors and areas in which our measurement of output are especially poor: services, information activities, health, and also the underground economy.

but at the end of the day he concluded that

> There have been many reasonable attempts to explain the productivity slowdown (. . . ), but no smoking gun has been found, and no single explanation appears to be able to account for all the facts, leaving the field in an unsettled state until this day.

[1]This has become a frequently repeated quote; its last occurrence in Gordon (2016, 543). Lipsey and Carlaw (2000) once concluded that "TFP is as much a measure of our ignorance as it is a measure of anything positive.", but what was meant by this "positive" remained unexplained!

[2]On Solow and the NBER approach, exemplified by Kendrick, see Kleiman et al. (1966).

Twenty years later, Banerjee and Duflo (2019) extensively reviewed the state of the art and drew basically the same conclusion.

It is instructive to consult case studies, for instance those concerning the Indian economy. Madsen et al. (2010) based their study on aggregate data over the period 1950–2005 and manufacturing firms data over the period 1993–2005. They concluded that there was "no robust long-run relationship between TFP and research activity" as well as that "TFP growth cannot be explained by growth in research activity." Ghosh and Parab (2021) continued by applying various endogenous growth models to aggregate TFP growth data over the period 1970–2017. Their results were mixed and, as far as positive, weak. There was a modest role for human capital and R&D, and a bit larger role for foreign direct investment (FDI).

Until the nineties, the research on productivity change typically made use of the concept of the 'representative firm' in combination with aggregate empirical data material provided by statistical agencies. The increased availability of longitudinal enterprise microdata sets has opened up many new, exciting research avenues. Researchers are now able to track large numbers of individual firms over time. This has led to a completely new area of research, with its own conferences and research centers.

## 1.2 The Neo-Classical Approach

A cornerstone of the neo-classical approach is the assumption that the technology governing the production unit or units under consideration (be it enterprises, industries, or economies) can be represented by a sufficiently neat primal or dual function of input and output quantities and/or prices and time. Thus, for example, let the technology that governs a certain unit's production process at time period $t$ be characterized by the cost function $C(w, y, t)$. This function provides the minimum cost for producing the output quantities $y$ (an $M$-dimensional vector) when the input prices are given by $w$ (an $N$-dimensional vector). Let the actual cost incurred by the production unit producing $y(t)$ be $w(t) \cdot x(t)$, where $x(t)$ denotes the vector of input quantities at period $t$ and $\cdot$ denotes the inner product.

One of the neo-classical key assumptions is that the production unit is a cost minimizer; that is, actual cost is assumed to be equal to minimum cost, or, in terms of the formulas just introduced:

$$w(t) \cdot x(t) = C(w(t), y(t), t). \tag{1.1}$$

Assuming that all functions are continuous, one can differentiate with respect to time $t$ to obtain growth rates. The left-hand side of expression (1.1) yields

$$\frac{d \ln w(t) \cdot x(t)}{dt} = \sum_n s_n(t) \frac{d \ln w_n(t)}{dt} + \sum_n s_n(t) \frac{d \ln x_n(t)}{dt}, \tag{1.2}$$

where the $s_n(t)$'s denote the actual cost shares of the inputs. The right-hand side of expression (1.1) yields

$$\frac{d \ln C(w(t), y(t), t)}{dt} = \sum_n \frac{\partial \ln C(.)}{\partial \ln w_n} \frac{d \ln w_n(t)}{dt} + \sum_m \frac{\partial \ln C(.)}{\partial \ln y_m} \frac{d \ln y_m(t)}{dt} + \frac{\partial \ln C(.)}{\partial t}. \tag{1.3}$$

The assumption of cost minimization combined with Shephard's Lemma leads to the well-known conclusion that

$$\frac{\partial \ln C(.)}{\partial \ln w_n} = s_n(t) \;\; (n = 1, \ldots, N); \tag{1.4}$$

that is, the partial derivatives occurring in the first term on the right-hand side of expression (1.3) are equal to the actual cost shares of the inputs.

The second assumption is that the output prices are proportional to marginal cost; that is

$$p_m(t) \propto \frac{\partial C(.)}{\partial y_m} \;\; (m = 1, \ldots, M). \tag{1.5}$$

Simple manipulations with this relation lead to the following expression:

$$\frac{\partial \ln C(.)}{\partial \ln y_m} = u_m(t) \sum_m \frac{\partial \ln C(.)}{\partial \ln y_m} \;\; (m = 1, \ldots, M), \tag{1.6}$$

where the $u_m(t)$'s are the actual revenue shares of the outputs.

The third assumption is that the technology exhibits constant returns to scale. This implies that

$$\sum_m \frac{\partial \ln C(.)}{\partial \ln y_m} = 1. \tag{1.7}$$

Thus, second and third assumption together imply that the partial derivatives occurring in the second term on the right-hand side of expression (1.3) are equal to the actual revenue shares of the outputs.

We can now put the various pieces together. Due to the cost minimization assumption the right-hand sides of expressions (1.2) and (1.3) are identically equal. Thus, substituting the two results just obtained, the following expression is obtained:

$$\sum_m u_m(t) \frac{d \ln y_m(t)}{dt} - \sum_n s_n(t) \frac{d \ln x_n(t)}{dt} = -\frac{\partial \ln C(.)}{\partial t}. \tag{1.8}$$

The left-hand side of this expression measures TFP change (in growth rate form) and is known as the Solow residual. The right-hand side is the minimum cost decrease associated with the mere passage of time. This is conventionally interpreted as a measure of technological change. Thus, under the stated assumptions, TFP change is equal to technological change.

The assumptions can be summarized as follows. First, the production unit is seen as cost efficient; that is, technically efficient—acting at the frontier of the current technology—, and allocatively efficient—the input quantities have the optimal mix. Second, the production unit is assumed to act in a competitive environment. Third, the technology is assumed to exhibit constant returns to scale. In the next chapter we will argue that as fourth assumption perfect foresight must be added.

## 1.3 The Contents of This Book

The measurement of productivity change (or difference) is usually based on models relying on strong assumptions such as competitive behaviour and constant returns to scale. Chapter 2 discusses the basics of (intertemporal) productivity measurement and shows that one can dispense with most if not all of the usual, neo-classical assumptions. Various models are reviewed and their relationships discussed. Throughout the chapter the equivalence of multiplicative and additive models, as well as the equivalence of productivity measurement and growth accounting, is highlighted. By virtue of their structural features, the various measurement models are applicable to individual establishments and aggregates such as industries, sectors, or economies.

Unlike labour productivity change, the measurement of total factor productivity change (or difference) crucially depends on the measurement and decomposition of capital input cost. Chapter 3 discusses the various measurement issues and shows that one can dispense with the usual neo-classical assumptions. There are several models and several alternatives for the 'rate of return'. However, the particular rate that must be used remains at the discretion of the researcher or the statistical agency.

Productivity change is generally measured in index form as ratio of an output quantity index over an input quantity index. Several statistical agencies publish quarterly as well as annual productivity index numbers or growth percentages, constructed from what appear to be basically the same sources. This raises the question whether, apart from measurement errors, consistency between quarterly and annual indices can be expected. Chapter 4 explores, from a theoretical perspective, the options for obtaining consistency between annual and quarterly (or more general: between period and subperiod) measures of productivity change.

An industry is usually an ensemble of individual firms (decision making units) which may or may not interact with each other. Similarly, an economy is an ensemble of industries. In National Accounts terms this is symbolized by the fact that the nominal value added produced by an industry or an economy is the simple sum of firm-, or industry-specific nominal value added. From this viewpoint it is natural to expect that there is a relation between (aggregate) industry or economy productivity and the (disaggregate) firm- or industry-specific productivities. This is the topic of Chaps. 5–9.

Chapter 5 considers the relation between (total factor) productivity measures for lower-level production units and aggregates thereof such as industries, sectors, or entire economies. In particular, this chapter contains a review of the so-called bottom-up approach, which takes an ensemble of individual production units, be it industries or enterprises, as the fundamental frame of reference. At the level of industries the various forms of shift-share analysis are reviewed. At the level of enterprises the additional features that must be taken into account are entry (birth) and exit (death) of production units.

The top-down approach is pursued in Chaps. 6–8. Chapter 6 is concerned with the relation between output and labour productivity measures for individual production units and for aggregates such as industries, sectors, or economies. In the framework of discrete time periods several useful, symmetric expressions are derived and confronted with results from the literature.

Chapter 7 proceeds by considering the relation between total factor productivity measures for individual production units and for aggregates such as industries, sectors, or economies. Though this topic has been treated in a number of influential publications, this chapter's distinctive feature is that all kinds of (neo-classical) structural and behavioural assumptions are avoided, such as assumptions about the existence of production frontiers with certain properties, or optimizing behaviour of the production units. In addition, the chapter treats dynamic ensembles of production units, characterized by entry and exit. Thus, a greater level of generality is achieved from which the earlier results follow as special cases.

In Chap. 7 three time-symmetric decompositions of aggregate value-added-based total factor productivity change were developed. In Chap. 8 a fourth decomposition will be developed. A notable difference with the previous chapter is that the development is cast in terms of levels rather than indices. Various aspects of this new decomposition will be discussed and links with decompositions found in the literature unveiled. It turns out that also here one can dispense with the usual neo-classical assumptions.

As said, productivity analysis is carried out at various levels of aggregation. In microdata studies the emphasis is on individual firms (or plants), whereas in sectoral studies it is on (groupings of) industries. Microdata researchers do not care too much about the interpretation of the weighted means of firm-specific productivities employed in their analyses. In Chap. 9 the consequences of this attitude are explored, based on a review of the literature.

However, a structurally similar phenomenon happens in sectoral studies, where the productivity change of industries is compared to each other and to the productivity change of some next-higher aggregate, which is usually the (measurable part of the) economy. Though there must be a relation between sectoral and economy-level measures, in most publications by statistical agencies and academic researchers this aspect is more or less neglected.

The point of departure of Chap. 9 is that aggregate productivity must be interpreted as productivity of the aggregate. It is shown that this implies restrictive relations between the productivity measure, the set of weights, and the type of mean employed.

The final chapter, Chap. 10, delves into the components of (total factor) productivity change. By making minimal assumptions about underlying technologies it appears that productivity change, here defined as output quantity change divided by input quantity change, can be seen as the combined result of (technical) efficiency change, technological change, a scale effect, and input- and output-mix effects. Given a certain functional form for the productivity index, the problem is then how to decompose such an index into factors corresponding to these five components. A basic insight offered in this chapter is that meaningful decompositions of productivity indices can only be obtained for indices that are transitive in the main variables. Using a unified approach, decompositions for Malmquist, Moorsteen-Bjurek, Lowe, and Cobb-Douglas productivity indices are obtained. A unique feature of this chapter is that all the decompositions are applied to the same dataset, a real-life panel of decision-making units, so that the extent of the differences between the various decompositions can be judged.

# Chapter 2
# A Framework Without Assumptions

## 2.1 Introduction

The methodological backing of productivity measurement and growth accounting usually goes like this. The (aggregate) production unit considered has an input side and an output side, and there is a production function that links output quantities to input quantities. This production function includes a time variable, and the partial derivative of the production function with respect to the time variable is called technological change (or, in some traditions, multi- or total factor productivity change). Further, it is assumed that the production unit acts in a competitive environment; that is, input and output prices are assumed as given. Next, it is assumed that the production unit acts in a profit maximizing manner (or, it is said to be 'in equilibrium'), and that the production function exhibits constant returns to scale. Under these assumptions it then appears that output quantity growth (defined as the output-share-weighted mean of the individual output quantity growth rates) is equal to input quantity growth (defined as the input-share-weighted mean of the individual input quantity growth rates) plus the rate of technological change (or, multi- or total factor productivity growth).

For the empirical implementation one then turns to National Accounts, census and/or survey data, in the form of nominal values and deflators (price indices). Of course, one cannot avoid dirty hands by making various imputations where direct observations failed or were impossible (as in the case of labour input of self-employed workers). In the case of capital inputs the prices, necessary for the computation of input shares, cannot be observed, but must be computed as unit user costs. The single degree of freedom that is here available, namely the rate of return, is used to ensure that the restriction implied by the assumption of constant returns to scale, namely that profit equals zero, is satisfied. This procedure is usually rationalized by the assumption of perfect foresight, which in this case means that the *ex post* calculated capital input prices can be assumed as *ex ante* given to the



B. M. Balk, *Productivity*, Contributions to Economics,
https://doi.org/10.1007/978-3-030-75448-8_2

production unit, so that they can be considered as exogenous data for the unit's profit maximization problem.

This account is, of course, somewhat stylized, since there occur many, smaller or larger, variations on this theme in the literature. An example was provided in Sect. 1.2. Recurring, however, are a number of so-called neo-classical assumptions: (1) a technology that exhibits constant returns to scale, (2) competitive input and output markets, (3) optimizing behaviour, and (4) perfect foresight. A fine example from academia is provided by Jorgenson et al. (2005, 23, 37), while the Sources and Methods publication of Statistics New Zealand (2006) shows that the neo-classical model has also deeply invaded official statistical agencies.[1] Another interesting example where neo-classical assumptions have invaded the measurement system is the World Productivity Database of the United Nations Industrial Development Organisation (UNIDO); see Isaksson (2009).

An interesting position is taken by the EU KLEMS Growth and Productivity Accounts project. Though in their main text Timmer et al. (2007) adhere to the Jorgenson, Ho and Stiroh framework, there is a curious footnote saying

> Under strict neo-classical assumptions, MFP [multifactor productivity] growth measures disembodied technological change. *In practice* [my emphasis], MFP is derived as a residual and includes a host of effects such as improvements in allocative and technical efficiency, changes in returns to scale and mark-ups as well as technological change proper. All these effects can be broadly summarized as "improvements in efficiency", as they improve the productivity with which inputs are being used in the production process. In addition, being a residual measure MFP growth also includes measurement errors and the effects from unmeasured output and inputs.

There are more examples of authors who exhibit similar concerns, without, however, feeling the need to adapt their conceptual framework.

For an official statistical agency, whose main task it is to provide statistics to many different users for many different purposes, it is discomforting to have such, strong and often empirically refuted, assumptions built into the methodological foundations of productivity and growth accounting statistics. This especially applies to the behavioural assumptions numbered 2, 3 and 4. There is ample evidence that, on average, markets are not precisely competitive; that producers' decisions frequently turn out to be less than optimal; and that managers almost invariably lack the magical feature of perfect foresight. Moreover, the environment in which production units operate is not so stable as the assumption of a fixed production function seems to claim.

Fortunately, it is possible to avoid making such assumptions. In a sense this book proposes to start where the usual story ends, namely at the empirical side.[2] For

[1]The neo-classical model figured already prominently in the 1979 report of the U. S. National Research Council's Panel to Review Productivity Statistics (Rees 1979). An overview of national and international practice is provided by the regularly updated *OECD Compendium of Productivity Indicators* (OECD 2019).

[2]There is another, minor, difference between the approach defended here and the usual story. The usual story runs in the framework of continuous time in which periods are of infinitesimal short

any production unit, the total factor productivity index is then *defined* as an output quantity index divided by an input quantity index. There are various options here, depending on what one sees as input and output, but the basic feature is that, given price and quantity (or value) data, this is simply a matter of index construction. There appear to be no behavioural assumptions involved, and this even applies—as will be demonstrated—to the construction of capital input prices. Surely, a number of imputations must be made (as in the case of the self-employed workers) and there is fairly large number of more or less defendable assumptions involved (for instance on the depreciation rates of capital assets), but this belongs to the daily bread and butter of economic statisticians.

Structural as well as behavioural assumptions enter the picture as soon as it comes to the *explanation* of productivity change. Then there are, depending on the initial level of aggregation, two main directions: (1) to explain productivity change at an aggregate level by productivity change and other factors operating at lower levels of aggregation; (2) to decompose productivity change into factors such as technological change, technical efficiency change, scale effects, input- and output-mix effects, and chance. In this case, to proceed with the analysis one cannot sidestep a technology model with certain specifications.

The contents of this chapter unfold as follows. Section 2.2 outlines the architecture of the basic, KLEMS-Y, input-output model, with its total and partial measures of productivity change. This section also links productivity measurement and growth accounting. Section 2.3 proceeds with the KL-VA and K-CF models. Four additional input-output models are briefly introduced in Sect. 2.4. This section also contains a comparison of all the models. Section 2.5 introduces the capital utilization rate. Section 2.6 reviews an empirical illustration. Section 2.7 concludes by discussing the main decomposition methods.

## 2.2 The Basic Input-Output Model

Let us consider a single production unit. This could be an establishment or plant, a firm, an industry, a sector, or even an entire economy. We will simply speak of a 'unit'. For the purpose of productivity measurement, such a unit is considered as a (consolidated) input-output system. What does this mean?

For the output side as well as for the input side there is some list of commodities (according to some classification scheme). A commodity is thereby defined as a set of closely related items (goods or services) which, for the purpose of analysis, can be considered as "equivalent", either in the static sense of their quantities being additive or in the dynamic sense of displaying equal relative price or quantity changes.

---

duration. When it then comes to implementation several approximations must be assumed. The approach in this book does not need this kind of assumptions either, because entirely based on accounting periods of finite duration, such as years.

Ideally, then, for any accounting period considered (*ex post*), say a year, each commodity comes with a value (in monetary terms) and a price and/or a quantity. If value and price are available, then the quantity is obtained by dividing the value by the price. If value and quantity are available, then the price is obtained by dividing the value by the quantity. If both price and quantity are available, then value is defined as price times quantity. In any case, for every commodity it must be so that value equals price times quantity, the magnitudes of which of course must pertain to the same accounting period. Technically speaking, the price concept used here is the unit value. At the output side, the prices must be those received by the unit, whereas at the input side, the prices must be those paid. Consolidation (also called net-sector approach) means that the unit does not deliver to itself. Put otherwise, all the intra-unit deliveries are netted out.

The inputs are customarily classified according to the KLEMS format. The letter K denotes the class of owned, reproducible capital assets. The commodities here are the asset-types, sub-classified by age category. Cohorts of assets are assumed to be available at the beginning of the accounting period and, in deteriorated form (due to ageing, wear and tear), still available at the end of the period. Investment during the period adds entities to these cohorts, while desinvestment, breakdown, or retirement remove entities. Examples include buildings and other structures, land, machinery, transport and ICT equipment, tools. Theory implies that quantities sought are just the quantities of all these cohorts of assets (together representing the productive capital stock), whereas the relevant prices are their unit user costs (per type-age combination), constructed from imputed interest rates, depreciation profiles, (anticipated) revaluations, and tax rates. The sum of quantities times prices then provides the capital input cost of a production unit.

The productive capital stock may be underutilized, which implies that not all the capital costs are incurred in actual production. See Schreyer (2001, section 5.6) for a general discussion of this issue. We return to this issue in Section 2.5.[3]

The letter L denotes the class of labour inputs; that is, all the types of work that are important to distinguish, cross-classified for instance according to educational attainment, gender, and experience (which is usually proxied by age categories). Quantities are measured as hours worked (or paid), and prices are the corresponding wage rates per hour. Where applicable, imputations must be made for the work carried out by self-employed persons. The sum of quantities times prices provides the labour input cost (or the labour bill, or labour compensation, as it is sometimes called).[4]

[3]A classic treatment in the neo-classical framework, based on the distinction between shortrun and longrun output or cost, is provided by Morrison Paul (1999). For a modern treatment (with keywords: representative firm, dynamic cost minimization, Cobb-Douglas production function with constant returns to scale, Hicks-neutral technological change) the reader is referred to Comin et al. (2020).

[4]The utilization rate of the labour input factors is assumed to be 1. Over- or underutilization from the point of view of jobs or persons is reflected in the wage rates. At the economy level, unutilized labour is called 'unemployment'.

The classes K and L concern so-called primary inputs. The letters E, M, and S denote three, disjunct classes of so-called intermediate inputs. First, E is the class of energy commodities consumed by a production unit: oil, gas, electricity, and water. Second, M is the class of all the (physical) materials consumed in the production process, which could be sub-classified into raw materials, semi-fabricates, and auxiliary products. Third, S is the class of all the business services which are consumed for maintaining the production process. This includes the services of leased capital assets and outsourced activities. Though it is not at all a trivial task to define precisely all the intermediate inputs and to classify them, it can safely be assumed that at the end of each accounting period there is a quantity and a price associated with each of those inputs.

Then, for each accounting period, production cost is defined as the sum of primary and intermediate input cost. Though this is usually not implemented at every aggregation level, there are good reasons to exclude R&D expenditure from production cost, the reason being that such expenditure is not related to the current production process but to a future one. Put otherwise, by performing R&D, production units try to shift the technology frontier. When it then comes to explaining productivity change, the non-exclusion of R&D expenditure might easily lead to a sort of double-counting error.[5]

At the output side, the letter Y denotes the class of commodities, goods and/or services, which are produced by the unit. Though in some industries, such as services industries or industries producing mainly unique goods, definitional problems are formidable, it can safely be assumed that for each accounting period there are data on quantities produced. For units operating on the market there are also prices. The sum of quantities times prices then provides the production revenue, and, apart from taxes on production, revenue minus cost yields profit.[6]

There is, of course, discussion possible about what to include or exclude at the input- and output-sides. We are here more or less tacitly assuming a broad production viewpoint, where for instance marketing services are included in the set S. A broader viewpoint would take into account sales and uses from inventories.[7]

Profit is an important financial performance measure. A somewhat less obvious, but equally useful, measure is ‘profitability’, defined as revenue *divided* by cost. Profitability gives, in monetary terms, the quantity of output per unit of input, and is thus a measure of return to aggregate input (and in some older literature called ‘return to the dollar’).

---

[5]There are a number of issues here, such as the separation of the R&D part of labour input, the precise definition of knowledge assets, and the distribution of R&D expenditures over the parts of multinational enterprises. See Diewert and Huang (2011) and De Haan and Haynes (2018).

[6]Sometimes zero profit is imposed by considering profit as the remuneration (price) for entrepreneurial activity (of which the quantity is set equal to one), and adding this to business services S. In terms of National Accounts profit equals gross operating surplus (GOS) minus the imputed income of self-employed persons and capital input cost.

[7]The inclusion of natural capital (including subsoil assets) in K is only meaningful if the production units are economies; see Brandt et al. (2017).

Monitoring the unit's performance over time is here understood to mean monitoring the development of its profit or its profitability. Both measures are, by nature, dependent on price and quantity changes, at the two sides of the unit. If there is (price) inflation and the unit's profit has increased then that mere fact does not necessarily mean that the unit has been performing better. Also, though general inflation does not influence the development of profitability, differential inflation does. If output prices have increased more than input prices then any increase of profitability does not necessarily imply that the unit has been performing better. Thus, for measuring the economic performance of the unit one wants to remove the effect of price changes, irrespective of whether those prices are within or beyond the unit's control.

Profit and profitability are different concepts. The first is a difference measure, the second is a ratio measure. Change of a variable through time, which will be our main focus, can also be measured by a difference or a ratio. It is important to realize that, apart from technical details—such as, that a ratio does not make sense if the variable changes sign or becomes equal to zero—, these two ways of measuring change are equivalent. Thus there appear to be a number of ways of mapping the same reality in numbers, but differing numbers do not necessarily imply differing realities.[8]

Profit change stripped of its price component will be called *real* profit change, and profitability change stripped of its price component will be called *real* profitability change.[9] Another name for real profit (-ability) change is (total factor) productivity change. Thus, productivity change can be measured as a ratio (namely as real profitability change) or as a difference (namely as real profit change). At the economy level, productivity change can be related to some measure of overall welfare change. A down-to-earth approach would use the National Accounts to establish a link between labour productivity change and real-income-per-capita change. A more sophisticated approach, using economic models and assumptions, was provided by Basu and Fernald (2002).

For a non-market unit the story must be told somewhat differently. For such a unit there are no output prices; hence, there is no revenue. Though there is cost, like for market units, there is no profit or profitability. National accountants usually resolve the problem here by *defining* the revenue of a non-market unit to be equal to its cost, thereby setting profit equal to 0 or profitability equal to 1.[10] But this leaves the problem that there is no natural way of splitting revenue change through time in real

[8]It is easy to see, for example, that increasing profit can occur simultaneously with decreasing profitability.

[9]Note that real change means nominal change deflated by some price index, not necessarily being a (headline) CPI. 'Stripping' is of course a vague term, and a more precise definition will be given later.

[10]This approach goes back to Hicks (1940).

and monetary components. This can only be carried out satisfactorily when there is some output quantity index that is independent from the input quantity index.[11]

It is useful to remind the reader that the notions of profit and profitability, though conceptually rather clear, are difficult to operationalize. One of the reasons is that cost includes the cost of owned capital assets, the measurement of which exhibits a substantial number of degrees of freedom, as we will see in the remainder of this chapter. Also, labour cost includes the cost of self-employed persons, for which wage rates and hours of work usually must be imputed. It will be clear that all these, and many other, uncertainties spill over to operational definitions of the profit and profitability concepts.

### 2.2.1 Notation

Let us now introduce some notation to define the various concepts we are going to use. As stated, at the output side we have $M$ items, each with their price (received) $p_m^t$ and quantity $y_m^t$, where $m = 1, \ldots, M$, and $t$ denotes an accounting period (say, a year). Similarly, at the input side we have $N$ items, each with their price (paid) $w_n^t$ and quantity $x_n^t$, where $n = 1, \ldots, N$. To avoid notational clutter, simple vector notation will be used throughout. All the prices and quantities are assumed to be positive, unless stated otherwise. The *ex post* accounting point-of-view will be used; that is, quantities and monetary values of the so-called flow variables (output and labour, energy, materials, services inputs) are realized values, complete knowledge of which becomes available after the accounting period has expired. Similarly, the cost of capital input is calculated *ex post*. This is consistent with official statistical practice.

The unit's revenue, that is, the value of its (gross) output, during the accounting period $t$ is defined as

$$R^t \equiv p^t \cdot y^t \equiv \sum_{m=1}^{M} p_m^t y_m^t, \tag{2.1}$$

whereas its (total) production cost is defined as

$$C^t \equiv w^t \cdot x^t \equiv \sum_{n=1}^{N} w_n^t x_n^t. \tag{2.2}$$

The unit's *profit* (including taxes on production) is then given by its revenue minus its cost; that is,

[11] More on this in Diewert (2018).

$$\Pi^t \equiv R^t - C^t = p^t \cdot y^t - w^t \cdot x^t. \tag{2.3}$$

The unit's (gross-output based) *profitability* (also including taxes on production) is defined as its revenue divided by its cost; that is,

$$\Upsilon^t \equiv R^t/C^t = p^t \cdot y^t/w^t \cdot x^t. \tag{2.4}$$

Given positive prices and quantities, it will always be the case that $R^t > 0$ and $C^t > 0$. Thus, profitability $\Upsilon^t$ is always positive, but profit $\Pi^t$ can be positive, negative, or zero. Profit depends on the size of the production unit, but profitability not. Thus, for comparing the performance of production units having different sizes profitability is the measure to use.

The next two performance concepts are margins. The first is the *profit-cost margin* of the production unit, defined as profit over cost,

$$\mu^t \equiv \frac{\Pi^t}{C^t}. \tag{2.5}$$

The relation between profit-cost margin and profitability is then given by

$$\mu^t = \Upsilon^t - 1; \tag{2.6}$$

that is, the profit-cost margin is the profitability expressed as a percentage (which is usually published as $\mu^t \times 100\%$). But what precisely does this mean?

To get a clue, consider first the single-output case; that is $M = 1$. Then the production unit's profitability reduces to

$$\Upsilon^t = p^t y^t/C^t = p^t/\left(C^t/y^t\right); \tag{2.7}$$

that is, price over cost per unit. Put otherwise, the profit-cost margin $\mu^t$ is then the *markup* of price over unit cost.

For the general, multi-output case, suppose that the cost can be allocated to the various outputs; that is, $C^t = \sum_{m=1}^{M} C_m^t$, where $C_m^t$ is the cost of producing $y_m^t$ units of output $m$ ($m = 1, \ldots, M$). Then the unit's profitability can be decomposed as

$$\Upsilon^t = \frac{\sum_{m=1}^{M} p_m^t y_m^t}{\sum_{m=1}^{M} C_m^t} = \sum_{m=1}^{M} \frac{C_m^t}{C^t} \frac{p_m^t}{C_m^t/y_m^t}. \tag{2.8}$$

Thus profitability is a cost-share weighted mean of output-specific price over unit-cost relatives. Put otherwise, the profit-cost margin is a cost-share weighted mean of output-specific markups,