

Springer Proceedings in Business and Economics

William H. Greene

Lynda Khalaf

Robin C. Sickles

Michael Veall

Marcel-Cristian Voia *Editors*

Productivity and Efficiency Analysis



Springer

Springer Proceedings in Business and Economics

More information about this series at <http://www.springer.com/series/11960>

William H. Greene • Lynda Khalaf
Robin C. Sickles • Michael Veall
Marcel-Cristian Voia
Editors

Productivity and Efficiency Analysis



Springer

Editors

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

Lynda Khalaf
Department of Economics
Carleton University
Ottawa, ON, Canada

Robin C. Sickles
Department of Economics
Rice University
Houston, TX, USA

Michael Veall
Department of Economics
McMaster University
Hamilton, ON, Canada

Marcel-Cristian Voia
Department of Economics
Carleton University
Ottawa, ON, Canada

ISSN 2198-7246

ISSN 2198-7254 (electronic)

Springer Proceedings in Business and Economics

ISBN 978-3-319-23227-0

ISBN 978-3-319-23228-7 (eBook)

DOI 10.1007/978-3-319-23228-7

Library of Congress Control Number: 2015956100

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Introduction

The volume comprises 17 chapters that deal with productivity measurement, productivity growth, dynamics of productivity change, measurement of labor productivity, measurement of technical efficiency at the sectoral level, frontier analysis, measurement of performance, industry instability, and spillover effects. The contributors to the volume are W. Erwin Diewert, Bert M. Balk, Subal C. Kumbhakar, Frank Asche, Kristin Roll, Ragnar Tveten, Loren W. Tauer, Jaepil Han, Deockhyun Ryu, Robin C. Sickles, Lynda Khalaf, Charles J. Saunders, German Cubas, Anson T.Y. Ho, Kim P. Huynh, David T. Jacho-Chàvez, A.S.J. Smith, J. Buckell, P. Wheat, R. Longo, Brian Murphy, Michael R. Veall, Yan Zhang, Yuri Ostrovsky, Robert J. Petrunia, Marcel C. Voia, Leonard Sabetti, Pat Adams, Weimin Wang, Alejandro Nin-Pratt, Roar Amundsen, Hilde Marit Kvile, Sourour Baccar, Mihailo Radoman, Jiaqi Hao, and Chenjun Shang.

The first chapter examines productivity decompositions at the sectoral level. The economy-wide labor productivity growth rate is thought to depend on sectoral labor productivity growth rates, real output price changes, and changes in sectoral labor input shares. A puzzle is that empirically, the real output price change effects, when aggregated across industries, have little explanatory power. The economy-wide TFP growth decomposition into sectoral explanatory factors depend on the sectoral TFP productivity growth rates, real output and input price changes, and changes in sectoral aggregate input shares. The puzzle with this decomposition is that empirically all of these price change effects and input share effects matter little when they were aggregated over sectors; only the sectoral TFP growth rates contributed significantly to overall TFP growth.

The second chapter 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, a review of the so-called bottom-up approach, which is an ensemble of individual production units, is considered. At the industry level, the various forms of shift-share analyses are reviewed.

The third chapter considers a revenue maximizing model and derives the revenue function from the transformation function where errors are parts of the outputs. The chapter adapts McElroy's additive general error model to the transformation

function with multiple inputs and multiple outputs and derives the revenue function. The error terms in the output supply functions, derived from the revenue function, inherit their randomness from the error terms in the outputs. As a second approach, the chapter uses a multiplicative general error model (MGEM), in the spirit of Kumbhakar and Tsionas (2011), with multiple outputs in which multiplicative errors are parts of the outputs. The MGEM is further generalized to accommodate output-oriented inefficiency. The translog revenue function with MGEM makes the intercept and the coefficients of the linear terms random (functions of the errors associated with outputs). Vessel level data for the Norwegian whitefish fisheries for the period 1995–2007 are used to showcase the application of the model. A standard (off-the-shelf) revenue function with output-oriented technical inefficiency is estimated and technical change and technical efficiency results are compared with the MGEM revenue function, which is estimated along with the revenue share equations. Although the means are found to be somewhat similar, patterns of technical change and technical efficiency are found to be quite different across these models.

Chapter 4 uses quantile regression to estimate production functions at various quantiles within a dairy farm production set, and marginal products and input substitutions are derived for each of the quantile production functions. Economic relationships vary in the interior of the production set compared to the frontier of the production set, with no discernible pattern through the production set. An implication is that the production response to inputs changes for inefficient firms in the interior of the production set may differ compared to efficient firms on the frontier of the production set. The results are for a specific dairy production dataset, so further analysis is warranted to determine what patterns exist with other empirical production sets.

Chapter 5 aims to investigate spillover effects of public capital stock in a production function model that accounts for spatial dependencies. Although there are a number of studies that estimate the output elasticity of public capital stock, they suffer from a failure to refine the output elasticity of public capital stock as well as to account for spillover effects of the public capital stock on the production efficiency when such spatial dependencies exist. A spatial autoregressive stochastic frontier model is employed and the authors analyze estimates with a time-varying spatial weights matrix. Using data for 21 OECD countries from 1960 to 2001, the chapter finds that spillover effects can be an important factor explaining variations in technical inefficiency across countries as well as in explaining the discrepancies among various levels of output elasticity of public capital stock in traditional production function approaches.

Chapter 6 considers a dynamic technical efficiency framework with non-Gaussian errors which suffer from the incidental parameter bias. Simulations show that an indirect inference estimation approach provides bias correction for the model and distribution parameters. The indirect confidence set inference method is size correct and exhibits good coverage properties even for asymmetric confidence regions. Bank cost data are examined under the proposed dynamic

technical efficiency framework with evidence that an MLE approach could provide misleading implications.

Chapter 7 studies labor productivity growth in Ecuador from 1998 to 2006 by using firm-level data from the annual survey of manufacturing and mining. This period is characterized by the economic crisis in 1999 and important economic reforms. During the crisis, there was a 2 % annual decrease in productivity in 1998–2000, but the recovery was strong with a 5 % annual productivity growth in 2002–2004. The productivity decompositions indicate that the main source of productivity growth came from firms with increasing productivity gaining market shares. Within-firm productivity decline was substantial during the crisis, but its growth was secondary in the post-crisis recovery. Firm entry and exit only had minor impacts on labor productivity. The distributional analysis further showed that labor productivity distribution increased in 2000–2002 and had remained at higher level for the rest of the sample period.

Chapter 8 looks at the source of inefficiency within health system organizational structures as a key aspect of performance measurement and management, which is of increasing importance to policy makers. The study uses a unique panel dataset to study the efficiency performance of pathology services in the National Health Service (NHS) in England for the first time. A dual-level stochastic frontier (DLSF) model (Smith and Wheat, 2012) to isolate the source of inefficiency at two vertically distinct organizational levels is used: an upper level of Strategic Health Authorities (SHAs) and a lower level of laboratories grouped within SHAs. A DLSF framework is developed in line with recent developments in the wider panel data literature to control for the influence unobserved heterogeneity, which is a key issue for healthcare performance analysis. Statistically significant variation in inefficiency performance at both organizational levels in pathology services is found. These measures are used to compute overall inefficiency for NHS pathology services, and corresponding savings estimates.

Chapter 9 looks at how productivity and growth may be affected by what are called “shortages” of specific types of workers. Canadian data are examined for evidence of a shortage of Information and Communication Technology (ICT) workers. Published vacancy and unemployment data are too coarse at the industry level. Accordingly, two types of administrative data are used to look for evidence of rising ICT employment and labor income, which might indicate a shortage. One dataset is available with little lag in cross section (from payroll records) and the other longitudinal dataset (based on tax filer data) is available with a 2-year lag. The results suggest that both data sources may be useful in this instance, with the longitudinal data used to check for compositional changes in the more current and timely cross-sectional data. Similar approaches may be available for other countries. These data sources provide at most mild evidence of a shortage of Canadian ICT workers in recent times.

Chapter 10 looks at the impact industry instability has on worker separations. Workers leave firms in one of two ways: (1) voluntarily by quitting or (2) involuntarily through firm layoffs. Using data drawn from the Longitudinal Worker File, a Canadian firm-worker matched employment database, the chapter distinguishes

between voluntary and involuntary separations using information on reasons for separations and assesses the impact industry shutdown rates have on worker separation rates, both voluntarily and involuntarily. Once controlling for various factors and potential selection bias, it is found that industry shutdown rates have a positive and significant effect on the overall separation, layoff and quit rates of workers. It is also found that industry instability has a much larger impact on layoff rates when comparing voluntary and involuntary separations.

Chapter 11 employs a growth-accounting approach to revisit past performance of agriculture in sub-Saharan Africa (SSA) and to analyze the relationship between the input mix used by SSA countries and productivity levels observed in the region. Findings show that improved technical efficiency has been the main driver of growth in recent years, benefiting poorer, low labor productivity countries. Countries with higher output and input per worker have benefited much more from technological progress than poorer countries, suggesting that technical change has done little to reduce the gap in labor productivity between countries. Results also show that the levels of input per worker used in SSA agriculture at present are extremely low and associated with less productive technologies, and that technical change has shifted the world technological frontier unevenly, increasing the distance between SSA countries and those countries with the “right” input mix.

Chapter 12 investigates the role of university knowledge spillovers in fostering innovative start-up firms, measured by R&D intensity, an important predictor of firm innovation and productivity. Annual data from the Kauffman Firm Survey of a representative cohort of US start-ups over the period 2004–2011 are used. By controlling for individual-firm characteristics and local factors, the chapter tests the effects of regional variation in R&D intensity of the higher education sector on start-up firms’ R&D expenditure decisions. Strong effects on both extensive and intensive margins of firm R&D expenditures are found. The results shed light on the role of entrepreneurs and new firm formation as a mechanism for innovation in universities as an important source of knowledge and technology transfer.

Chapter 13 presents a growth-accounting framework in which subsoil mineral and energy resources are recognized as natural capital input into the production process in two ways. Firstly, the income attributable to subsoil resources, or resource rent, is estimated as a surplus value after all extraction costs and normal returns on produced capital have been accounted for. The value of a resource reserve is then estimated as the present value of the future resource rents generated from the efficient extraction of the reserve. Secondly, with extraction as the observed service flows of natural capital, multifactor productivity growth and sources of economic growth can be reassessed by updating income shares of all inputs and then by estimating the contribution to growth coming from changes in the value of natural capital input. The empirical results on the Canadian oil and gas extraction show that adding natural capital increases the annual multifactor productivity growth in the oil and gas sector from –2.3 to –1.5 % over the 1981–2009 period. During the same period, the annual real value-added growth in this industry was 2.3 %, of which about 0.4 percentage points or 16 % comes from natural capital.

Chapter 14 is a presentation and discussion of issues that arise in the practical application of a regulatory benchmarking model. It describes the regulatory benchmarking model for electricity distribution companies in Norway, and focuses on how different choices influence different incentives for the companies. These choices cover methodology, modeling assumptions, and variables, but also how the benchmarking results are applied in the regulatory model. The benchmarking model is only one part of the regulatory model for setting revenue caps. The discussion shows some of the trade-offs that have to be considered in this process, and sheds some light on why regulators may deviate from optimal textbook solutions.

Chapter 15 evaluates the capacity of the translog cost share model to approximate the producer's true demand system and introduces two nonlinear functional forms, which have been achieved by altering and extending the standard quadratic logarithmic translog model. The extensions have additional desirable approximation properties with respect to output and time variables, and thus allow more flexible treatments of non-homothetic technologies and non-neutral technical change than those provided by the standard translog. The performances of the three models are assessed (1) on theoretical ground, by the size of the domain of regularity, (2) on their ability to provide plausible estimates of the economic and technological indicators being measured, and finally (3) on their reliability in fitting input shares, input–output ratios, and unit cost. The most important finding is that the standard model exhibits some weakness in fitting. The authors show via a series of experiments that those shortcomings are due to a lack of flexibility of the logarithmic model. The estimation results obtained with the new extended model are promising.

Chapter 16 examines the impact of policy changes on player productivity at the top level of European football, with a particular focus on the English Premier League. Contest theory motivates the prediction that post-Bosman entrants will be more productive and consequently have a higher probability of earning/retaining a first-team spot in top European leagues. To test these predictions, data were collected on all players that entered the English Premier League in 4-year windows around the Bosman ruling. Nonparametric techniques, specifically Regression Discontinuity Design, were applied to test for sharp jumps in player productivity measures around the Bosman ruling. The results display discontinuity in player productivity measures, suggesting that post-Bosman entrants tend to be more productive than pre-Bosman entrants.

Chapter 17 provides new methods to robustify productivity growth measurement by utilizing various economic theories explaining economic growth and productivity and the econometric model generated by that particular theory. The World Productivity Database from the UNIDO is utilized to analyze productivity during the period 1960–2010 for OECD countries. The focus is on three competing models from the stochastic frontier literature, Cornwell, Schmidt, and Sickles (1990), Kumbhakar (1990), and Battese and Coelli (1992) to estimate productivity growth and its decomposition into technical change and efficiency change and utilize methods due to Hansen (2010) to construct optimal weights in order to model average the results from these three approaches.

Contents

1	Decompositions of Productivity Growth into Sectoral Effects: Some Puzzles Explained.....	1
	W. Erwin Diewert	
2	The Dynamics of Productivity Change: A Review of the Bottom-Up Approach	15
	Bert M. Balk	
3	A General Error Revenue Function Model with Technical Inefficiency: An Application to Norwegian Fishing Trawler	51
	Subal C. Kumbhakar, Frank Asche, Kristin Roll, and Ragnar Tveteras	
4	Production Response in the Interior of the Production Set	71
	Loren W. Tauer	
5	Spillover Effects of Public Capital Stock Using Spatial Frontier Analyses: A First Look at the Data	83
	Jaepil Han, Deockhyun Ryu, and Robin C. Sickles	
6	Dynamic Technical Efficiency	99
	Lynda Khalaf and Charles J. Saunders	
7	Analysing Labour Productivity in Ecuador	109
	German Cubas, Anson T. Y. Ho, Kim P. Huynh, and David T. Jacho-Chávez	
8	Hierarchical Performance and Unobservable Heterogeneity in Health: A Dual-Level Efficiency Approach Applied to NHS Pathology in England.....	119
	A.S.J. Smith, J. Buckell, P. Wheat and R. Longo	
9	Is There Evidence of ICT Skill Shortages in Canadian Taxfiler Data?	145
	Brian Murphy, Michael R. Veall and Yan Zhang	

10	Worker Separations and Industry Instability	161
	Kim P. Huynh, Yuri Ostrovsky, Robert J. Petrunia, and Marcel-Cristian Voia	
11	Inputs, Productivity and Agricultural Growth in Sub-Saharan Africa	175
	Alejandro Nin-Pratt	
12	University Knowledge Spillovers and Innovative Startup Firms.....	203
	Leonard Sabetti	
13	Accounting for Natural Capital in Productivity of the Mining and Oil and Gas Sector	211
	Pat Adams and Weimin Wang	
14	Balancing Incentives: The Development and Application of a Regulatory Benchmarking Model	233
	Roar Amundsveen and Hilde Marit Kvile	
15	Limitations of the Approximation Capabilities of the Translog Model: Implications for Energy Demand and Technical Change Analysis	249
	Sourour Baccar	
16	Bosman Ruling Implications on Player Productivity in the English Premier League	291
	Mihailo Radoman	
17	Productivity Measurement, Model Averaging, and World Trends in Growth and Inequality.....	305
	Robin C. Sickles, Jiaqi Hao, and Chenjun Shang	
Index		325

Chapter 1

Decompositions of Productivity Growth into Sectoral Effects: Some Puzzles Explained

W. Erwin Diewert

Abstract An earlier paper by Diewert (*J Prod Anal* 43(3):367–387, 2015) provided some new decompositions of economy wide labour productivity growth and Total Factor Productivity (TFP) growth into sectoral effects. The economy wide labour productivity growth rate turned out to depend on the sectoral labour productivity growth rates, real output price changes and changes in sectoral labour input shares. A puzzle is that empirically, the real output price change effects, when aggregated across industries, did not matter much. The economy wide TFP growth decomposition into sectoral explanatory factors turned out to depend on the sectoral TFP productivity growth rates, real output and input price changes and changes in sectoral aggregate input shares. The puzzle with this decomposition is that empirically all of these price change effects and input share effects did not matter much when they were aggregated over sectors; only the sectoral TFP growth rates contributed significantly to overall TFP growth. The present paper explains these puzzles.

Keywords Total Factor Productivity • Labour productivity • Index numbers • Sectoral contributions to growth

JEL code: C43, C82, D24

1.1 Introduction

Many analysts want to decompose aggregate labour productivity growth (or aggregate Multifactor growth) into sectoral effects so that the industry sources of productivity growth can be determined. However, it turns out that measures of economy wide productivity change cannot be obtained as a simple weighted sum

W.E. Diewert (✉)

School of Economics, University of British Columbia, Vancouver, BC, Canada V6T 1Z1

The School of Economics, University of New South Wales, Sydney, NSW, Australia
e-mail: erwin.diewert@ubc.ca

of the corresponding industry measures; Denison (1962) showed that changes in the allocation of resources across the industries also played an important role in contributing to aggregate labour productivity change. However, the Denison decomposition of aggregate labour productivity change into explanatory effects ignored the role of changes in industry output prices. In an important paper, Tang and Wang (2004, p. 426) extended the Denison decomposition to take into account changes in real output prices in their decomposition of economy wide labour productivity into explanatory contribution effects. However, Tang and Wang combined the effects of changes in real output prices with the effects of changes in input shares and so Diewert (2015) extended their contribution to provide a decomposition of aggregate labour productivity growth into separate contribution terms due to sectoral productivity growth, changes in input shares and changes in real output prices. In Sect. 1.2 below, we present Diewert's decomposition of aggregate labour productivity growth into explanatory sectoral factors. Diewert (2015) further extended the Tang and Wang methodology in order to provide a decomposition of economy wide Total Factor Productivity (or Multifactor Productivity) growth into industry explanatory factors.¹ Section 1.3 presents this generalization.

However, in Diewert's (2015) empirical examples using Australian data which illustrated his new decompositions, he found that many of the industry explanatory factors, when summed over industries, were very small. In Sects. 1.4 and 1.5 below, we will explain why these somewhat puzzling results hold. Section 1.4 explains the labour productivity puzzles while Sect. 1.5 explains the Total Factor Productivity growth decomposition puzzles.

1.2 Diewert's Aggregate Labour Productivity Growth Decomposition

Let there be N sectors or industries in the economy.² Suppose that for period $t = 0, 1$, the *output* (or real value added or volume) of sector n is Y_n^t with corresponding period t *price* P_n^t ³ and *labour input* L_n^t for $n = 1, \dots, N$. We assume that these labour inputs can be added across sectors and that the *economy wide labour input* in period t is L^t defined as follows:

¹Other recent papers dealing with decompositions of labour productivity include Diewert (2010), de Avillez (2012), Dumagan (2013), Balk (2014a, b), Dumagan and Balk (2014) and Reinsdorf (2014).

²The material in this section and the following one is taken from Diewert (2015).

³These industry real output aggregates Y_n^t and the corresponding prices P_n^t are *indexes* of the underlying micro net outputs produced by industry n . The exact functional form for these indexes does not matter for our analysis (with some mild restrictions to be noted later) but we assume the indexes satisfy the property that for each t and n , $P_n^t Y_n^t$ equals the industry n nominal value added for period t .

$$L^t \equiv \sum_{n=1}^N L_n^t; \quad t = 0, 1. \quad (1.1)$$

Industry n labour productivity in period t , X_n^t , is defined as industry n real output divided by industry n labour input:

$$X_n^t \equiv Y_n^t / L_n^t; \quad t = 0, 1; \quad n = 1, \dots, N. \quad (1.2)$$

It is not entirely clear how aggregate labour productivity should be defined since the outputs produced by the various industries are measured in heterogeneous units, which are in general, not comparable. Thus it is necessary to weight these heterogeneous outputs by their prices, sum the resulting period t values and then divide by an appropriate *output price index*, say P^t for period t , in order to make the economy wide nominal value of aggregate output comparable in real terms across periods.⁴ Thus with an appropriate choice for the aggregate output price index P^t , the period t *economy wide labour productivity*, X^t , is defined as follows⁵:

$$X^t \equiv \sum_{n=1}^N P_n^t Y_n^t / P^t L^t = \sum_{n=1}^N \left(P_n^t / P^t \right) Y_n^t / L^t = \sum_{n=1}^N P_n^t Y_n^t / L^t; \quad t = 0, 1 \quad (1.3)$$

where the *period t industry n real output price*, p_n^t , is defined as the industry t output price P_n^t , divided by the aggregate output price index for period t , P^t ; i.e., we have the following definitions⁶:

$$P_n^t \equiv P_n^t / P^t, \quad n = 1, \dots, N; t = 0, 1. \quad (1.4)$$

⁴An economy wide sequence of real value added output levels Y^t (and the corresponding value added output price levels P^t) is generally defined by aggregating individual industry outputs (and intermediate inputs entered with negative signs) into economy wide output levels, using bilateral Laspeyres, Paasche, Fisher (1922) or Törnqvist indexes. These indexes use the industry P_n^t and Y_n^t for the two periods being compared and either fixed base or chained aggregate price and quantities (the P^t and Y^t) are generated for each period in the sample. If Fisher indexes are used as the basic formula, then for each year t , the product of the Fisher aggregate price and quantity levels, $P^t Y^t$, will equal the sum of industry price times volumes for year t , $\sum_{n=1}^N P_n^t Y_n^t$, which in turn is equal to market sector nominal value added for year t . Thus the year t aggregate quantity levels Y^t are equal to $\sum_{n=1}^N P_n^t Y_n^t / P^t$ and so (1.3) can be rewritten as $X^t = Y^t / L^t$.

⁵This follows the methodological approach taken by Tang and Wang (2004, p. 425). As noted in the previous footnote, the aggregate output price index P^t can be formed by applying an index number formula to the industry output prices (or value added deflators) for period t , (P_1^t, \dots, P_N^t) , and the corresponding real output quantities (or industry real value added estimates) for period t , (Y_1^t, \dots, Y_N^t) . The application of chained superlative indexes would be appropriate in this context but again, the exact form of index does not matter for our analysis as long as $P^t Y^t$ equals period t aggregate nominal value added.

⁶These definitions follow those of Tang and Wang (2004, p. 425).

Using definitions (1.2) and (1.3), it is possible to relate the period t aggregate productivity level X^t to the industry productivity levels X_n^t as follows⁷:

$$\begin{aligned} X^t &\equiv \sum_{n=1}^N P_n^t Y_n^t / P^t L^t \\ &= \sum_{n=1}^N P_n^t [Y_n^t / L_n^t] [L_n^t / L^t] \\ &= \sum_{n=1}^N P_n^t s_{Ln}^t X_n^t \quad \text{using definitions (2)} \end{aligned} \quad (1.5)$$

where the *share of labour used by industry n in period t*, s_{Ln}^t , is defined in the obvious way as follows:

$$s_{Ln}^t \equiv L_n^t / L^t; \quad n = 1, \dots, N; \quad t = 0, 1. \quad (1.6)$$

Thus aggregate labour productivity for the economy in period t is a weighted sum of the sectoral labour productivities where the weight for industry n is p_n^t , the real output price for industry n in period t , times s_{Ln}^t , the share of labour used by industry n in period t .

Up to this point, the above analysis follows that of Tang and Wang (2004, pp. 425–426) but now we follow Diewert's (2015) approach.⁸

First, Diewert defined the *value added or output share of industry n in total value added for period t*, s_{Yn}^t , as follows:

$$\begin{aligned} s_{Yn}^t &\equiv P_n^t Y_n^t / \sum_{i=1}^N P_i^t Y_i^t \quad t = 0, 1; \quad n = 1, \dots, N \\ &= P_n^t Y_n^t / \sum_{i=1}^N p_i^t Y_i^t \quad \text{using definitions (4).} \end{aligned} \quad (1.7)$$

Diewert noted that the product of the sector n real output price times its labour share in period t , $p_n^t s_{Ln}^t$, with the sector n labour productivity in period t , X_n^t , equals the following expression:

$$\begin{aligned} P_n^t s_{Ln}^t X_n^t &= P_n^t [L_n^t / L^t] [Y_n^t / L_n^t]; \quad t = 0, 1; \quad n = 1, \dots, N \\ &= P_n^t Y_n^t / L^t. \end{aligned} \quad (1.8)$$

⁷Equation (1.5) corresponds to (1.2) in Tang and Wang (2004, p. 426).

⁸Tang and Wang (2004, pp. 425–426) combined the effects of the real price for industry n for period t , p_n^t , with the industry n labour share s_{Ln}^t for period t by defining the relative size of industry n in period t , s_n^t , as the product of p_n^t and s_{Ln}^t ; i.e., they defined the industry n weight in period t as $s_n^t \equiv P_n^t s_{Ln}^t$. They then rewrote (1.5), $X^t = \sum_{n=1}^N P_n^t s_{Ln}^t X_n^t$, as $X^t = \sum_{n=1}^N s_n^t X_n^t$. Thus their analysis of the effects of the changes in the weights s_n^t did not isolate the separate effects of changes in industry real output prices and industry labour input shares.

Using definition (1.3) and (1.5), aggregate labour productivity growth (plus 1) going from period 0 to 1, X^1/X^0 , is equal to:

$$\begin{aligned} X^1/X^0 &= \sum_{n=1}^N p_n^1 s_{Ln}^1 X_n^1 / \sum_{n=1}^N p_n^0 s_{Ln}^0 X_n^0 \\ &= \sum_{n=1}^N [p_n^1/p_n^0] [s_{Ln}^1/s_{Ln}^0] [X_n^1/X_n^0] \\ &\quad [p_n^0 Y_n^0/L^0] / \sum_{i=1}^N [p_i^0 Y_i^0/L^0] \text{ using (8)} \\ &= \sum_{n=1}^N [p_n^1/p_n^0] [s_{Ln}^1/s_{Ln}^0] [X_n^1/X_n^0] s_{Yn}^0 \text{ using definitions (7).} \end{aligned} \quad (1.9)$$

Thus overall economy wide labour productivity growth, X^1/X^0 , is an output (or value added) share weighted average of three *growth factors* associated with industry n. The three growth factors are:

- X_n^1/X_n^0 , (one plus) the rate of growth in the labour productivity of industry n;
- s_{Ln}^1/s_{Ln}^0 , (one plus) the rate of growth in the share of labour being utilized by industry n and
- $p_n^1/p_n^0 = [P_n^1/P_n^0]/[P^1/P^0]$ which is (one plus) the rate of growth in the real output price of industry n.

Thus in looking at the contribution of industry n to overall (one plus) labour productivity growth, start out with a straightforward share weighted contribution factor, $s_{Yn}^0[X_n^1/X_n^0]$, which is the period 0 output or value added share of industry n in period 0, s_{Yn}^0 , times the industry n rate of labour productivity growth (plus one), X_n^1/X_n^0 . This straightforward contribution factor for industry n will be augmented if real output price growth for industry n is positive (if p_n^1/p_n^0 is greater than one) and if the share of labour used by industry n is growing (if s_{Ln}^1/s_{Ln}^0 is greater than one). The decomposition of overall labour productivity growth given by the last line of (1.9) seems to be intuitively reasonable and fairly simple as opposed to the decomposition obtained by Tang and Wang (2004, p. 426) which does not separately distinguish the effects of real output price change from changes in the industry's labour share.

The puzzle with the above decomposition (1.9) is that empirically, it appears that the effects of real output price change, when aggregated over industries, are insignificant; i.e., Diewert (2015) found that for the case of Australia, the following decomposition of aggregate labour productivity growth provided a close approximation to the exact decomposition (1.9):

$$X^1/X^0 \approx \sum_{n=1}^N [s_{Ln}^1/s_{Ln}^0][X_n^1/X_n^0] s_{Yn}^0. \quad (1.10)$$

Note that the real output price change augmentation factors (the p_n^1/p_n^0) are not present on the right hand side of (1.10). Thus for the case of Australia, only the

labour share augmentation factors s_{Ln}^1/s_{Ln}^0 and the industry labour productivity growth rates X_n^1/X_n^0 proved to be significant determinants of overall labour productivity growth.⁹ In Sect. 1.4 below, we will show why this result holds in general.

1.3 Diewert's Aggregate Multifactor Productivity Growth Decomposition

As in the previous section, let there be N sectors or industries in the economy and again suppose that for period $t = 0, 1$, the *output* (or real value added or volume) of sector n is Y_n^t with corresponding period t *price* P_n^t . However, it is now assumed that each sector uses many inputs and index number techniques are used to form industry input aggregates Z_n^t with corresponding aggregate industry input prices W_n^t for $n = 1, \dots, N$ and $t = 0, 1$.¹⁰

Industry n Total Factor Productivity (TFP) in period t, X_n^t , is defined as industry n real output Y_n^t divided by industry n real input Z_n^t :

$$X_n^t \equiv Y_n^t/Z_n^t; \quad t = 0, 1; \quad n = 1, \dots, N. \quad (1.11)$$

As in Sect. 1.2, *economy wide real output in period t*, Y^t , is defined as total value added divided by the economy wide output price index P^t . Thus for each period t , the following identity holds¹¹:

$$Y^t = \sum_{n=1}^N P_n^t Y_n^t / P^t = \sum_{n=1}^N P_n^t Y_n^t; \quad t = 0, 1 \quad (1.12)$$

where the *period t industry n real output price* is defined as $P_n^t \equiv P_n^t/P^t$ for $n = 1, \dots, N$ and $t = 0, 1$.

⁹This does not mean that the industry real output price augmentation factors [p_n^1/p_n^0] were all close to one (they were not for Diewert's Australian data); it just means that when aggregating over industries, the factors which were greater than one are balanced by factors less than one so that the effects of real output price change cancel out when aggregating over industries.

¹⁰These industry input aggregates Z_n^t and the corresponding price indexes W_n^t are *indexes* of the underlying micro inputs utilized by industry n . The exact functional form for these indexes does not matter for our analysis but it is assumed that the indexes satisfy the property that for each t and n , $W_n^t Z_n^t$ equals the industry n input cost for period t .

¹¹As in Sect. 1.2, $P_n^t Y_n^t$ is nominal industry n value added in period t . Industry n period t real value Y_n^t added is defined as period t nominal industry n value added deflated by the industry n value added price index P_n^t . It need not be the case that $P_n^t Y_n^t$ is equal to $W_n^t Z_n^t$; i.e., it is not necessary that value added equal input cost for each industry for each time period.

Economy wide real input in an analogous manner. Thus an *economy wide input price index* for period t , W^t , is formed in one of two ways¹²:

- By aggregating over all industry micro economic input prices using microeconomic input quantities as weights to form W^t (single stage aggregation of inputs) or
- By aggregating the industry aggregate input prices W_n^t (with corresponding input quantities or volumes Z_n^t) into the aggregate period t input price index W^t using an appropriate index number formula (two stage aggregation of inputs).

Economy wide real input in period t, Z^t , is defined as economy wide input cost divided by the economy wide input price index W^t ¹³:

$$Z^t = \sum_{n=1}^N W_n^t Z_n^t / W^t = \sum_{n=1}^N w_n^t Z_n^t; \quad t = 0, 1 \quad (1.13)$$

where the *period t industry n real input price* is defined as:

$$w_n^t \equiv W_n^t / W^t; \quad n = 1, \dots, N \text{ and } t = 0, 1. \quad (1.14)$$

The *economy wide level of TFP (or MFP) in period t*, X^t , is defined as aggregate real output divided by aggregate real input:

$$X^t \equiv Y^t / Z^t; \quad t = 0, 1. \quad (1.15)$$

Denote the output share of industry n in period t , s_{Yn}^t , by (1.7) again and define the *input share of industry n in economy wide cost in period t*, s_{Zn}^t , as follows:

$$\begin{aligned} s_{Zn}^t &\equiv W_n^t Z_n^t / \sum_{i=1}^N W_i^t Z_i^t \quad t = 0, 1; \quad n = 1, \dots, N \\ &= w_n^t Z_n^t / \sum_{i=1}^N w_i^t Z_i^t \end{aligned} \quad (1.16)$$

where the second in (1.16) follows from the definitions $w_n^t \equiv W_n^t / W^t$.

Substitute (1.12) and (1.13) into definition (1.15) and the following expression for the economy wide level of TFP in period t is obtained:

¹²In either case, it is assumed that the product of the total economy input price index for period t , W^t , with the corresponding aggregate input quantity or volume index, Z^t , is equal to total economy nominal input cost. If Laspeyres or Paasche price indexes are used throughout, then the two stage and single stage input aggregates will coincide. If superlative indexes are used throughout, then the two stage and single stage input aggregates will approximate each other closely using annual data; see Diewert (1978).

¹³Note that $W^t Z^t$ equals period t total economy input cost for each t .

$$\begin{aligned}
X^t &= \sum_{n=1}^N P_n^t Y_n^t / \sum_{i=1}^N w_i^t Z_i^t \quad t = 0, 1 \\
&= \sum_{n=1}^N P_n^t (Y_n^t / Z_n^t) Z_n^t / \sum_{i=1}^N w_i^t Z_i^t \\
&= \sum_{n=1}^N (P_n^t / w_n^t) X_n^t w_n^t Z_n^t / \sum_{i=1}^N w_i^t Z_i^t \quad \text{using (11)} \\
&= \sum_{n=1}^N (P_n^t / w_n^t) X_n^t s_{Zn}^t \quad \text{using (16).}
\end{aligned} \tag{1.17}$$

Using (1.17), aggregate TFP growth (plus 1) going from period 0 to 1, X^1/X^0 , is equal to:

$$\begin{aligned}
X^1/X^0 &= \sum_{n=1}^N (p_n^1 / w_n^0) X_n^1 s_{Zn}^1 / \sum_{i=1}^N (p_i^0 / w_i^0) X_i^0 s_{Zi}^0 \\
&= \sum_{n=1}^N (p_n^1 / p_n^0) (w_n^0 / w_n^1) (X_n^1 / X_n^0) (s_{Zn}^1 / s_{Zn}^0) (p_n^0 / w_n^0) \\
&\quad X_n^0 s_{Zn}^0 / \sum_{i=1}^N (p_i^0 / w_i^0) X_i^0 s_{Zi}^0 \\
&= \sum_{n=1}^N s_{Yn}^0 (p_n^1 / p_n^0) (w_n^0 / w_n^1) (s_{Zn}^1 / s_{Zn}^0) (X_n^1 / X_n^0).
\end{aligned} \tag{1.18}$$

The last in (1.18) follows from the following equations for $n = 1, \dots, N$:

$$\begin{aligned}
(p_n^0 / w_n^0) X_n^0 s_{Zn}^0 &= (p_n^0 / w_n^0) (Y_n^0 / Z_n^0) (w_n^0 Z_n^0 / \sum_{i=1}^N w_i^0 Z_i^0) \\
&\quad \text{using (11) and (16)} \\
&= p_n^0 Y_n^0 / \sum_{i=1}^N w_i^0 Z_i^0.
\end{aligned} \tag{1.19}$$

Thus one plus economy wide TFP growth, X^1/X^0 , is equal to an output share weighted average (with the base period weights s_{Yn}^0) of one plus the industry TFP growth rates, times an augmentation factor, which is the product $(p_n^1/p_n^0)(w_n^0/w_n^1)(s_{Zn}^1/s_{Zn}^0)$. Thus formula (1.18) is very similar to the previous labour productivity growth formula (1.9) except that now there is an additional multiplicative augmentation factor, which is w_n^0/w_n^1 , the reciprocal of one plus the rate of growth of real input prices for sector n.

Diewert's empirical results for Australia indicated that while individual industry terms for any of the above contribution terms on the right hand side of (1.18) can be quite significant, the aggregate effects of the real output and input price augmentation factors as well as the labour input share factors were close to zero; i.e., for Diewert's Australian data, it appeared that the following decomposition of aggregate Multifactor productivity growth provided a close approximation to the exact decomposition given by (1.18):

$$X^1/X^0 \approx \sum_{n=1}^N s_{Yn}^0 [X_n^1/X_n^0]. \tag{1.20}$$

Note that the real output price change augmentation factors (the p_n^1/p_n^0), the reciprocal real input price change augmentation factors (the w_n^0/w_n^1) and the input share augmentation factors (the s_{Zn}^1/s_{Zn}^0) are not present on the right hand side of (1.20). Thus for the case of Australia, only the industry multifactor productivity growth rates X_n^1/X_n^0 proved to be significant determinants of overall Multifactor

productivity growth.¹⁴ In Sect. 1.5 below, we will show why this somewhat puzzling result holds in general.¹⁵

1.4 The Labour Productivity Growth Decomposition Puzzle Explained

We want to show that the exact identity (1.9) can be approximated by the right hand side of (1.10) in Sect. 1.2 above.

The starting point in the derivation of the approximate identity (1.10) is to note that if the bilateral index number formula that is used to form the period t economy wide output aggregates and the corresponding aggregate output price levels is the direct or implicit Laspeyres, Paasche, Fisher, Törnqvist or any other known superlative index number formula, then it can be shown that the output price levels generated by Cobb-Douglas index number formula will approximate the price levels corresponding to any of the above formulae to the first order around an equal price and quantity point.¹⁶ Recall that the aggregate output price index levels for periods 0 and 1 were defined as P^0 and P^1 . Thus using this above approximation result, it can be seen that to the accuracy of a first order Taylor series approximation, the following approximate equality will hold:

$$\ln \left[P^1 / P^0 \right] \approx \sum_{n=1}^N s_{Y_n}^0 \ln \left[p_n^{-1} / p_n^0 \right]. \quad (1.21)$$

Subtracting $\ln[P^1/P^0]$ from both sides of (1.21) and making use of definitions (1.4) for the industry real output prices p_n^t , it can be seen that (1.21) becomes the following approximate equality:

$$0 \approx \sum_{n=1}^N s_{Y_n}^0 \ln \left[p_n^{-1} / p_n^0 \right]. \quad (1.22)$$

From (1.9), we have X^1/X^0 equal to the weighted arithmetic mean of the N numbers, $[p_n^{-1}/p_n^0][s_{L_n}^{-1}/s_{L_n}^0][X_n^{-1}/X_n^0]$ with weights $s_{Y_n}^0$. Approximate this weighted arith-

¹⁴This does not mean that the industry augmentation factors p_n^{-1}/p_n^0 , w_n^0/w_n^{-1} and $s_{Z_n}^{-1}/s_{Z_n}^0$ were all close to one (they were not for Diewert's Australian data); it just means that when aggregating over industries, the factors which were greater than one are balanced by factors less than one so that the effects of real output and input price changes and of industry cost share changes cancel out when aggregating over industries.

¹⁵In order to derive the approximate MFP growth decomposition (1.18), we will require one additional assumption; namely that the value of inputs equals the value of outputs less intermediate inputs for each industry in period 0.

¹⁶See Diewert (1978) for a proof of this result. Note that this approximation result also holds for indexes built up in two stages.

metic mean by the corresponding weighted geometric mean¹⁷ and take logarithms of the resulting approximate equality. We obtain the following approximate equality:

$$\begin{aligned}\ln [X^1/X^0] &\approx \sum_{n=1}^N s_{Yn}^0 \ln \left[p_n^1/p_n^0 \right] + \sum_{n=1}^N s_{Yn}^0 \ln \left[s_{Ln}^{-1}/s_{Ln}^0 \right] \\ &\quad + \sum_{n=1}^N s_{Yn}^0 \ln \left[X_n^1/X_n^0 \right] \\ &\approx \sum_{n=1}^N s_{Yn}^0 \ln \left[s_{Ln}^{-1}/s_{Ln}^0 \right] + \sum_{n=1}^N s_{Yn}^0 \ln \left[X_n^1/X_n^0 \right]\end{aligned}\quad (1.23)$$

where we have used (1.22) to establish the second approximate equality. Now exponentiate both sides of (1.23) and approximate the geometric mean on the right hand side by the corresponding arithmetic mean and we obtain the following approximate equality:

$$X^1/X^0 \approx \sum_{n=1}^N s_{Yn}^0 \left[s_{Ln}^{-1}/s_{Ln}^0 \right] \left[X_n^1/X_n^0 \right] \quad (1.24)$$

which is the approximate equality (1.10). Thus to the accuracy of a first order Taylor series approximation, the industry real output augmentation factors p_n^1/p_n^0 , in aggregate, do not contribute to overall labour productivity growth.

1.5 The Multifactor Productivity Growth Decomposition Puzzle Explained

In order to derive the approximate aggregate MFP growth decomposition defined by (1.20) in Sect. 1.3, it is necessary to make another assumption. The extra assumption is that for each industry, the value of primary inputs is exactly equal to the value of outputs less intermediate input costs for the base period. This extra assumption implies that the industry output shares will equal the industry primary input shares so that the following equalities will be satisfied¹⁸:

$$s_{Yn}^0 = s_{Zn}^0; \quad n = 1, \dots, N. \quad (1.25)$$

We also assume that the bilateral index number formula that is used to form the period t economy wide input aggregates is the Laspeyres, Paasche, Fisher, Törnqvist or any other known superlative index number formula. Again, it can be shown that

¹⁷It is easy to show that a weighted geometric mean of N positive numbers will approximate the corresponding weighted arithmetic mean of the same numbers to the first order around a point where the numbers are all equal.

¹⁸Most national statistical agencies that compute Multifactor Productivity use balancing rates of return in their user costs in order to make the value of inputs equal to the value of outputs; i.e., for most statistical agencies that compute TFP growth, this assumption will be satisfied.

the corresponding Cobb-Douglas index number formula for the price index will approximate any of the above formulae to the first order around an equal price and quantity point. Recall that the aggregate input price index levels for periods 0 and 1 were defined as W^0 and W^1 . Thus using this above approximation result, it can be seen that to the accuracy of a first order Taylor series approximation, the following approximate equality will hold:

$$\ln[W^1/W^0] \approx \sum_{n=1}^N s_{Zn}^0 \ln \left[W_n^1 / W_n^0 \right]. \quad (1.26)$$

Subtracting $\ln[W^1/W^0]$ from both sides of (1.26) and making use of definitions (1.14) for the industry real input prices w_n^t , it can be seen that (1.26) becomes the following approximate equality:

$$\begin{aligned} 0 &\approx \sum_{n=1}^N s_{Zn}^0 \ln \left[w_n^1 / w_n^0 \right] \\ &= \sum_{n=1}^N s_{Yn}^0 \ln \left[w_n^1 / w_n^0 \right] \text{ using assumptions (25)} \\ &= -\sum_{n=1}^N s_{Yn}^0 \ln \left[w_n^0 / w_n^1 \right] \\ &= \sum_{n=1}^N s_{Yn}^0 \ln \left[w_n^0 / w_n^1 \right] \end{aligned} \quad (1.27)$$

where the last equality follows from the fact that $-0 = 0$.

From (1.18), we have X^1/X^0 equal to the weighted arithmetic mean of the N numbers, $(p_n^1/p_n^0)(w_n^0/w_n^1)(s_{Zn}^1/s_{Zn}^0)(X_n^1/X_n^0)$ with weights s_{Yn}^0 . Approximate this weighted arithmetic mean by the corresponding weighted geometric mean and take logarithms of the resulting approximate equality. We obtain the following approximate equality:

$$\begin{aligned} \ln[X^1/X^0] &\approx \sum_{n=1}^N s_{Yn}^0 \ln \left[p_n^1 / p_n^0 \right] + \sum_{n=1}^N s_{Yn}^0 \ln \left[w_n^0 / w_n^1 \right] \\ &\quad + \sum_{n=1}^N s_{Yn}^0 \ln \left[s_{Zn}^1 / s_{Zn}^0 \right] + \sum_{n=1}^N s_{Yn}^0 \ln \left[X_n^1 / X_n^0 \right] \\ &\approx \sum_{n=1}^N s_{Yn}^0 \ln \left[s_{Zn}^1 / s_{Zn}^0 \right] + \sum_{n=1}^N s_{Yn}^0 \ln \left[X_n^1 / X_n^0 \right] \quad (1.28) \end{aligned}$$

using (22) and (27).

Exponentiate both sides of (1.28) and on the right hand side of the resulting approximate identity, we obtain the product of the weighted geometric means of the s_{Zn}^1/s_{Zn}^0 and of the X_n^1/X_n^0 using the output share weights s_{Yn}^0 . Approximate both weighted geometric means by their corresponding weighted arithmetic means and we obtain the following approximate equality:

$$\begin{aligned}
X^1/X^0 &\approx \left\{ \sum_{n=1}^N s_{Yn}^0 [s_{Zn}^{-1}/s_{Zn}^0] \right\} \left\{ \sum_{n=1}^N s_{Yn}^0 [X_n^{-1}/X_n^0] \right\} \\
&= \left\{ \sum_{n=1}^N s_{Yn}^0 [s_{Yn}^{-1}/s_{Yn}^0] \right\} \left\{ \sum_{n=1}^N s_{Yn}^0 [X_n^{-1}/X_n^0] \right\} \quad \text{using (25)} \\
&= \left\{ \sum_{n=1}^N s_{Yn}^{-1} \right\} \left\{ \sum_{n=1}^N s_{Yn}^0 [X_n^{-1}/X_n^0] \right\} \\
&= \left\{ \sum_{n=1}^N s_{Yn}^0 [X_n^{-1}/X_n^0] \right\}
\end{aligned} \tag{1.29}$$

where the last equation follows since $\sum_{n=1}^N s_{Yn}^{-1} = 1$; i.e., the period 1 industry shares of total economy value added sum up to unity in period 1.¹⁹ Thus we have derived the very simple approximate TFP growth decomposition into industry explanatory factors defined by (1.20) in Sect. 1.3.

1.6 Conclusion

We have explained the rather puzzling empirical results obtained by Diewert (2015) in his analysis of sectoral factors that contributed to economy wide labour productivity growth and to economy wide TFP or MFP growth. The simplification of the general decomposition formulae given by (1.9) for labour productivity and by (1.18) for TFP into (1.10) and (1.20) respectively are due to the definitions of the productivity concepts as an index number of aggregate output growth divided by either aggregate labour growth or an index number of aggregate input growth. The use of index numbers to define aggregate output and input leads to a cancellation of output and input price effects in the aggregate decompositions. In the case of the TFP decomposition, the assumption that value added equals input cost for each industry in each period leads to a cancellation of input allocation effects in the aggregate TFP decomposition. Thus if analysts want to focus on industry explanatory factors that do not cancel out in the aggregate, the simplified approximate formulae defined by (1.10) and (1.20) are recommended. On the other hand, if analysts want to focus on the individual contributions of each industry to productivity growth (whether the effects cancel out in the aggregate or not), then the more complex exact formulae defined by (1.9) and (1.18) are recommended.

Acknowledgements The author thanks Bert Balk for helpful comments and gratefully acknowledges the financial support of the SSHRC of Canada.

¹⁹Note the role of assumptions (1.25) in the derivation of (1.29). The assumption that industry primary input costs equal the corresponding industry net revenues is what enables us to deduce the unimportance of industry cost share reallocations in the aggregate when decomposing aggregate TFP growth into industry explanatory factors. Note that in the case of Labour Productivity growth, we did *not* assume that industry labour input shares were equal to industry net revenue shares and so changes in labour input shares did play an important role in the approximate decomposition of aggregate Labour Productivity growth given by (1.24) in the previous section.

References

- Balk BM (2014a) Dissecting aggregate labour and output productivity change. *J Prod Anal* 42: 35–43
- Balk BM (2014b) Measuring and relating aggregate and subaggregate total factor productivity change without neoclassical assumptions. *Statistica Neerlandica* 69(1):21–48
- de Avillez R (2012) Sectoral contributions to labour productivity growth in Canada: does the choice of decomposition formula matter? *Prod Monit Number* 24(Fall):97–117
- Denison EF (1962) The sources of economic growth in the United States and the alternatives before us. Committee for Economic Development, New York
- Diewert WE (1978) Superlative index numbers and consistency in aggregation. *Econometrica* 46:883–900
- Diewert WE (2010) On the Tang and Wang decomposition of labour productivity into sectoral effects. In: Diewert WE, Balk BM, Fixler D, Fox KJ, Nakamura AO (eds) Price and productivity measurement, vol 6. Index Number Theory, Trafford Press, Victoria, pp 67–76
- Diewert WE (2015) Decompositions of productivity growth into sectoral effects. *J Prod Anal* 43(3):367–387
- Dumagan JC (2013) A generalized exactly additive decomposition of aggregate labor productivity growth. *Rev Income Wealth* 59:157–168
- Dumagan JC, Balk BM (2014) Dissecting aggregate output and labour productivity change: a postscript on the role of relative prices. Rotterdam School of Management, Erasmus University, Rotterdam
- Fisher I (1922) The making of index numbers. Houghton-Mifflin, Boston
- Reinsdorf M (2015) Measuring industry contributions to labour productivity change: A new formula and a chained Fisher framework, *International Productivity Monitor* 28:3–26
- Tang J, Wang W (2004) Sources of aggregate labour productivity growth in Canada and the United States. *Can J Econ* 37:421–444

Chapter 2

The Dynamics of Productivity Change: A Review of the Bottom-Up Approach

Bert M. Balk

Abstract This paper 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 paper 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.

Keywords Producer • Productivity • Aggregation • Decomposition • Shift-share analysis • Bottom-up approach • Index number theory

JEL code: C43, O47

2.1 Introduction

In a previous article (Balk 2010) I considered the measurement of productivity change for a single, consolidated production unit.¹ The present paper continues by studying an ensemble of such units. The classical form is a so-called sectoral shift-share analysis. The starting point of such an analysis is an ensemble of industries,

An extended version of this paper is available at SSRN: <http://ssrn.com/abstract=2585452>.

¹“Consolidated” means that intra-unit deliveries are netted out. In some parts of the literature this is called “sectoral”. At the economy level, “sectoral” output reduces to GDP plus imports, and “sectoral” intermediate input to imports.

B.M. Balk (✉)

Rotterdam School of Management, Erasmus University, Rotterdam, The Netherlands

e-mail: bbalk@rsm.nl

according to some industrial classification (such as ISIC or NAICE), at some level of detail. An industry is a set of enterprises² engaged in the same or similar kind of activities. In the case of productivity analysis the ensemble is usually confined to industries for which independent measurement of input and output is available. Such an ensemble goes by different names: business sector, market sector, commercial sector, or simply measurable sector. Data are published and/or provided by official statistical agencies.

Let us, by way of example, consider labour productivity, in particular value-added based labour productivity. The output of industry k at period t is then measured as real value added RVA^{kt} ; that is, nominal value added VA^{kt} (= revenue minus intermediate inputs cost) deflated by a suitable, ideally industry-specific, price index. Real value added is treated as ‘quantity’ of a single commodity, that may or may not be added across the production units belonging to the ensemble studied, and over time. At the input side there is usually given some simple measure of labour input, such as total number of hours worked L^{kt} ; rougher measures being persons employed or full time equivalents employed. Then labour productivity of industry k at period t is defined as RVA^{kt}/L^{kt} .

In the ensemble the industries are of course not equally important, thus some weights reflecting relative importance, θ^{kt} , adding up to 1, are necessary. In the literature there is some discussion as to the precise nature of these weights. Should the weights reflect (nominal) value-added shares $VA^{kt}/\sum_k VA^{kt}$? or real value-added shares $RVA^{kt}/\sum_k RVA^{kt}$? or labour input shares $L^{kt}/\sum_k L^{kt}$? We return to this discussion later on.

Aggregate labour productivity at period t is then defined as a weighted mean, either arithmetic $\sum_k \theta^{kt} RVA^{kt}/L^{kt}$ or geometric $\prod_k (RVA^{kt}/L^{kt})^{\theta^{kt}}$, and the focus of interest is the development of such a mean over time.³ There are clearly two main factors here, shifting importance and shifting productivity, and their interaction. The usual product of a shift-share analysis is a table which provides detailed decomposition results by industry and time periods compared. Special interest can be directed thereby to industries which are ICT-intensive, at the input and/or the output side; industries which are particularly open to external trade; industries which are (heavily) regulated; *etcetera*.

Things become only slightly more complicated when value-added based *total factor productivity* is considered. At the input side one now needs per industry and time period nominal capital and labour cost as well as one or more suitable deflators. The outcome is real primary input, X_{KL}^{kt} , which can be treated as ‘quantity’ of another single commodity. Total factor productivity of industry k at period t is then defined

²There is no unequivocal naming here. So, instead of enterprises one also speaks of firms, establishments, plants, or kind-of-activity units. The minimum requirement is that realistic annual profit/loss accounts can be compiled.

³Curiously, the literature neglects the harmonic mean $(\sum_k \theta^{kt} (RVA^{kt}/L^{kt})^{-1})^{-1}$; but, as will be shown in the extended version of this paper, there are conditions under which this mean materializes as the natural one.

as RVA^{kt}/X_{KL}^{kt} . The issue of the precise nature of the weights gets some additional complexity, since we now also could contemplate the use of nominal or real cost shares to measure the importance of the various industries.

More complications arise when one wants to base the analysis on *gross-output* based total factor productivity. For the output side of the industries one then needs nominal revenue as well as suitable, industry-specific deflators. For the input side one needs nominal primary and intermediate inputs cost together with suitable deflators. The question of which weights to use is aggravated by the fact that industries deliver to each other, so that part of one industry's output becomes part of another industry's input. Improper weighting can then easily lead to double-counting of productivity effects.

Since the early 1990s an increasing number of statistical agencies made (longitudinal) microdata of enterprises available for research. Economists could now focus their research at production units at the lowest level of aggregation and dispense with the age-old concept of the 'representative firm' that had guided so much theoretical development. At the firm or enterprise level one usually has access to nominal data about output revenue and input cost detailed to various categories, in addition to data about employment and some aspects of financial behaviour. Lowest level quantity data are usually not available, so that industry-level deflators must be used. Also, at the enterprise level the information available is generally insufficient to construct firm-specific capital stock data. Notwithstanding such practical restrictions, microdata research has spawned and is still spawning lots of interesting results. A landmark contribution, including a survey of older results, is Foster et al. (2001). Good surveys were provided by Bartelsman and Doms (2000) and, more recently, Syverson (2011). Recent examples of research are collected in a special issue on firm dynamics of the journal *Structural Change and Economic Dynamics* 23 (2012), 325–402.

Of course, dynamics at the enterprise level is much more impressive than at the industry level, no matter how fine-grained. Thought-provoking features are the growth, decline, birth, and death of production units. Split-ups as well as mergers and acquisitions occur all over the place. All this is exacerbated by the fact that the annual microdata sets are generally coming from (unbalanced, rotating) samples, which implies that any superficial analysis of given datasets is likely to draw inaccurate conclusions.

This paper contains a review and discussion of the so-called *bottom-up* approach, which takes an ensemble of individual production units as the fundamental frame of reference. The theory developed here can be applied to a variety of situations, such as (1) a large company consisting of a number of subsidiaries, (2) an industry consisting of a number of enterprises, or (3) an economy or, more precisely, the 'measurable' part of an economy consisting of a number of industries.

The *top-down* approach is the subject of three other papers, namely Balk (2014, 2015) and Dumagan and Balk (2015). The connection between the two approaches, bottom-up and top-down, is discussed in the extended version of the present paper.

What may the reader expect from this review? Sections 2.2 and 2.3 describe the scenery: a set of production units with their accounting relations, undergoing temporal change. Section 2.4 defines the various measurement devices, in particular productivity indices, levels, and their links. The second half of this section is devoted to a discussion of the gap between theory and practice; that is, what to do when not all the data wanted are accessible? And what are the consequences of approximations?

Aggregate productivity change can be measured in different ways. First, as the development through time of arithmetic means of production-unit-specific productivity levels. Section 2.5 reviews the various decompositions proposed in the extant literature, and concludes with a provisional evaluation. Next, Sect. 2.6 briefly discusses the alternatives which emerge when arithmetic means are replaced by geometric or harmonic means. Section 2.7 discusses the monotonicity “problem”, revolving around the so-called Fox “paradox”: an increase of all the individual productivities not necessarily leads to an increase of aggregate productivity. It is argued that this is not a paradox at all but an essential feature of aggregation. Section 2.8 delves into the foundations of the much-used Olley-Pakes decomposition and distinguishes between valid and fallacious use.

In the bottom-up approach aggregate productivity is some weighted mean of individual, production-unit-specific productivities. There is clearly a lot of choice here: in the productivity measure, in the weights of the units, and in the type of mean. Section 2.9 formulates the problem; the actual connection, however, between the bottom-up and top-down approaches is discussed in the extended version of this paper. Section 2.10 concludes with a summary of the main lessons.

2.2 Accounting Identities

We consider an ensemble (or set) \mathcal{K}^t of consolidated production units,⁴ operating during a certain time period t in a certain country or region. For each unit the KLEMS-Y *ex post* accounting identity in nominal values (or, in current prices) reads

$$C_{KL}^{kt} + C_{EMS}^{kt} + \Pi^{kt} = R^{kt} \quad (k \in \mathcal{K}^t), \quad (2.1)$$

where C_{KL}^{kt} denotes the primary input cost, C_{EMS}^{kt} the intermediate inputs cost, R^{kt} the revenue, and Π^{kt} the profit (defined as remainder). Intermediate inputs cost (on energy, materials, and business services) and revenue concern generally tradeable commodities. It is presupposed that there is some agreed-on commodity classification, such that C_{EMS}^{kt} and R^{kt} can be written as sums of quantities times (unit) prices of these commodities. Of course, for any production unit most of these

⁴In terms of variables to be defined below, consolidation means that $C_{EMS}^{kkt} = R^{kkt} = 0$.

quantities will be zero. It is also presupposed that output prices are available from a market or else can be imputed. Taxes on production are supposed to be allocated to the K and L classes.

The commodities in the capital class K concern owned tangible and intangible assets, organized according to industry, type, and age class. Each production unit uses certain quantities of those assets, and the configuration of assets used is in general unique for the unit. Thus, again, for any production unit most of the asset cells are empty. Prices are defined as unit user costs and, hence, capital input cost C_K^{kt} is a sum of prices times quantities.

Finally, the commodities in the labour class L concern detailed types of labour. Though any production unit employs specific persons with certain capabilities, it is usually their hours of work that count. Corresponding prices are hourly wages. Like the capital assets, the persons employed by a certain production unit are unique for that unit. It is presupposed that, wherever necessary, imputations have been made for self-employed workers. Henceforth, labour input cost C_L^{kt} is a sum of prices times quantities.

Total primary input cost is the sum of capital and labour input cost, $C_{KL}^{kt} = C_K^{kt} + C_L^{kt}$. Profit Π^{kt} is the balancing item and thus may be positive, negative, or zero.

The KL-VA accounting identity then reads

$$C_{KL}^{kt} + \Pi^{kt} = R^{kt} - C_{EMS}^{kt} \equiv VA^{kt} \quad (k \in \mathcal{K}^t), \quad (2.2)$$

where VA^{kt} denotes value added, defined as revenue minus intermediate inputs cost. In this paper it will always be assumed that $VA^{kt} > 0$.

We now consider whether the ensemble of production units \mathcal{K}^t can be considered as a consolidated production unit. Though aggregation basically is addition, adding-up the KLEMS-Y relations over all the units would imply double-counting because of deliveries between units. To see this, it is useful to split intermediate input cost and revenue into two parts, respectively concerning units belonging to the ensemble \mathcal{K}^t and units belonging to the rest of the world. Thus,

$$C_{EMS}^{kt} = \sum_{k' \in \mathcal{K}^t, k' \neq k} C_{EMS}^{k'kt} + C_{EMS}^{ekt}, \quad (2.3)$$

where $C_{EMS}^{k'kt}$ is the cost of the intermediate inputs purchased by unit k from unit k' , and C_{EMS}^{ekt} is the cost of the intermediate inputs purchased by unit k from the world beyond the ensemble \mathcal{K} . Similarly,

$$R^{kt} = \sum_{k' \in \mathcal{K}^t, k' \neq k} R^{kk't} + R^{ket}, \quad (2.4)$$

where $R^{kk't}$ is the revenue obtained by unit k from delivering to unit k' , and R^{ket} is the revenue obtained by unit k from delivering to units outside of \mathcal{K}^t . Adding up the KLEMS-Y relations (2.1) then delivers

$$\begin{aligned} \sum_{k \in \mathcal{K}^t} C_{KL}^{kt} + \sum_{k \in \mathcal{K}^t} \sum_{k' \in \mathcal{K}, k' \neq k} C_{EMS}^{k'kt} + \sum_{k \in \mathcal{K}^t} C_{EMS}^{ekt} + \sum_{k \in \mathcal{K}^t} \Pi^{kt} = \\ \sum_{k \in \mathcal{K}^t} \sum_{k' \in \mathcal{K}^t, k' \neq k} R^{kk't} + \sum_{k \in \mathcal{K}^t} R^{ket}. \end{aligned} \quad (2.5)$$

If for all the tradeable commodities output prices are identical to input prices (which is ensured by National Accounting conventions), then the two intra- \mathcal{K}^t -trade terms cancel, and the foregoing expression reduces to

$$\sum_{k \in \mathcal{K}^t} C_{KL}^{kt} + \sum_{k \in \mathcal{K}^t} C_{EMS}^{ekt} + \sum_{k \in \mathcal{K}^t} \Pi^{kt} = \sum_{k \in \mathcal{K}^t} R^{ket}. \quad (2.6)$$

Recall that capital assets and hours worked are unique for each production unit, which implies that primary input cost may simply be added over the units, without any fear for double-counting. Thus expression (2.6) is the KLEMS-Y accounting relation for the ensemble \mathcal{K}^t , considered as a consolidated production unit. The corresponding KL-VA relation is then

$$\sum_{k \in \mathcal{K}^t} C_{KL}^{kt} + \sum_{k \in \mathcal{K}^t} \Pi^{kt} = \sum_{k \in \mathcal{K}^t} R^{ket} - \sum_{k \in \mathcal{K}^t} C_{EMS}^{ekt}, \quad (2.7)$$

which can be written as

$$C_{KL}^{\mathcal{K}^t} + \Pi^{\mathcal{K}^t} = R^{\mathcal{K}^t} - C_{EMS}^{\mathcal{K}^t} \equiv VA^{\mathcal{K}^t}, \quad (2.8)$$

where $C_{KL}^{\mathcal{K}^t} \equiv \sum_{k \in \mathcal{K}^t} C_{KL}^{kt}$, $\Pi^{\mathcal{K}^t} \equiv \sum_{k \in \mathcal{K}^t} \Pi^{kt}$, $R^{\mathcal{K}^t} \equiv \sum_{k \in \mathcal{K}^t} R^{ket}$, and $C_{EMS}^{\mathcal{K}^t} \equiv \sum_{k \in \mathcal{K}^t} C_{EMS}^{ekt}$. One verifies immediately that

$$VA^{\mathcal{K}^t} = \sum_{k \in \mathcal{K}^t} VA^{kt}. \quad (2.9)$$

The similarity between expressions (2.2) and (2.8), together with the additive relation between all the elements, is the reason why the KL-VA production model is the natural starting point for studying the relation between individual and aggregate measures of productivity change. We will see however that the bottom-up approach basically neglects this framework.

2.3 Continuing, Entering, and Exiting Production Units

As indicated in the previous section the superscript t denotes a time period, the usual unit of measurement being a year. Though data may be available over a longer time span, any comparison is concerned with only two periods: an earlier period

0 (also called base period), and a later period 1 (also called comparison period). These periods may or may not be adjacent. When the production units are industries, then the ensemble \mathcal{K}^0 will usually be the same as \mathcal{K}^1 . But when the production units studied are enterprises, this will in general not hold, and we must distinguish between continuing, exiting, and entering production units. In particular,

$$\mathcal{K}^0 = \mathcal{C}^{01} \cup \mathcal{X}^0 \quad (2.10)$$

$$\mathcal{K}^1 = \mathcal{C}^{01} \cup \mathcal{N}^1, \quad (2.11)$$

where \mathcal{C}^{01} denotes the set of continuing units (that is, units active in both periods), \mathcal{X}^0 the set of exiting units (active in the base period only), and \mathcal{N}^1 the set of entering units (active in the comparison period only). The sets \mathcal{C}^{01} and \mathcal{X}^0 are disjunct, as are \mathcal{C}^{01} and \mathcal{N}^1 .

It is important to observe that in any application the distinction between continuing, entering, and exiting production units depends on the length of the time periods being compared, and on the time span between these periods.

Of course, when the production units studied form a balanced panel, then the sets \mathcal{X}^0 and \mathcal{N}^1 are empty. The same holds for the case where the production units are industries. These two situations will in the sequel be considered as specific cases.

The theory developed in the remainder of this paper is cast in the language of intertemporal comparisons. By redefining 0 and 1 as countries or regions, and conditioning on a certain time period, the following can also be applied to cross-sectional comparisons. There is one big difference, however. Apart from mergers, acquisitions and the like, enterprises have a certain perseverance and can be observed through time. But a certain enterprise cannot exist at the same time in two countries or regions. Hence, in cross-sectional comparisons the lowest-level production units can only be industries, and ‘entering’ or ‘exiting’ units correspond to industries existing in only one of the two countries or regions which are compared.

2.4 Productivity Indices and Levels

As explained in Sect. 2.2, the various components of the accounting identity (2.1) are nominal values, that is, sums of prices times quantities. We are primarily interested in their development through time, as measured by ratios. It is assumed that all the detailed price and quantity data, underlying the values, are accessible. This is, of course, the ideal situation, which in practice is not likely to occur. Nevertheless, for conceptual reasons it is good to use this as our starting point. More mundane situations, deviating to a higher or lesser degree from the ideal, will then be considered later.

2.4.1 Indices

Using index number theory, each nominal value ratio can be decomposed as a product of two components, one capturing the price effect and the other capturing the quantity effect. Thus, let there be price and quantity indices such that for any two periods t and t' the following relations hold:

$$C_{KL}^{kt}/C_{KL}^{kt'} = P_{KL}^k(t, t')Q_{KL}^k(t, t') \quad (2.12)$$

$$C_{EMS}^{kt}/C_{EMS}^{kt'} = P_{EMS}^k(t, t')Q_{EMS}^k(t, t') \quad (2.13)$$

$$R^{kt}/R^{kt'} = P_R^k(t, t')Q_R^k(t, t'). \quad (2.14)$$

Capital cost and labour cost are components of primary input cost, thus it can also be assumed that there are functions such that

$$C_K^{kt}/C_K^{kt'} = P_K^k(t, t')Q_K^k(t, t') \quad (2.15)$$

$$C_L^{kt}/C_L^{kt'} = P_L^k(t, t')Q_L^k(t, t'). \quad (2.16)$$

We are using here the shorthand notation introduced in the earlier article (Balk 2010). All these price and quantity indices are supposed to be, appropriately dimensioned, functions of the prices and quantities at the two periods that play a role in the value ratios; e.g. $P_L^k(t, t')$ is a labour price index for production unit k , based on all the types of labour distinguished, comparing hourly wages at the two periods t and t' , conditional on hours worked at these periods. These functions are supposed to satisfy some basic axioms ensuring proper behaviour, and, dependent on the time span between t and t' , may be direct or chained indices (see Balk 2008). There may or may not exist functional relations between the overall index $P_{KL}^k(t, t')$ and the subindices $P_K^k(t, t')$ and $P_L^k(t, t')$ (or, equivalently, between the overall index $Q_{KL}^k(t, t')$ and the subindices $Q_K^k(t, t')$ and $Q_L^k(t, t')$).

The construction of price and quantity indices for value added was discussed in Balk (2010, Appendix B). Thus there are also functions such that

$$VA^{kt}/VA^{kt'} = P_{VA}^k(t, t')Q_{VA}^k(t, t') \quad (2.17)$$

Formally, the relations (2.12), (2.13), (2.14), (2.16) and (2.17) mean that the Product Test is satisfied. Notice that it is not required that all the functional forms of the price and quantity indices be the same. However, the Product Test in combination with the axioms rules out a number of possibilities.

We recall some definitions. The *value-added based total factor productivity index* for period 1 relative to period 0 was defined by Balk (2010) as

$$ITFPROM^k_{VA}(1, 0) \equiv \frac{Q_{VA}^k(1, 0)}{Q_{KL}^k(1, 0)}. \quad (2.18)$$

This index measures the ‘quantity’ change component of value added relative to the quantity change of all the primary inputs. The two main primary input components are capital and labour; both deserve separate attention.

The *value-added based capital productivity index* for period 1 relative to period 0 is defined as

$$IKPROD^k_{VA}(1, 0) \equiv \frac{Q_{VA}^k(1, 0)}{Q_K^k(1, 0)}. \quad (2.19)$$

This index measures the ‘quantity’ change component of value added relative to the quantity change of capital input.

Similarly, the *value-added based labour productivity index* for period 1 relative to period 0 is defined as

$$ILPROD^k_{VA}(1, 0) \equiv \frac{Q_{VA}^k(1, 0)}{Q_L^k(1, 0)}. \quad (2.20)$$

This index measures the ‘quantity’ change component of value added relative to the quantity change of labour input. Recall that the labour quantity index $Q_L^k(t, t')$ is here defined as an index acting on the prices and quantities of all the types of labour that are being distinguished.

Suppose now that the units of measurement of the various types of labour are in some sense the same; that is, the quantities of all the labour types are measured in hours, or in full-time equivalent jobs, or in some other common unit. Then it makes sense to define the total labour quantity of production unit k at period t as

$$L^{kt} \equiv \sum_{n \in L} x_n^{kt}, \quad (2.21)$$

and to use the ratio $L^{kt}/L^{kt'}$ as quantity index. Formally, this is a Dutot or simple sum quantity index. The ratio of a genuine labour quantity index, i.e. an index based on types of labour, $Q_L^k(t, t')$, and the simple sum labour quantity index $L^{kt}/L^{kt'}$ is an index of *labour quality (or composition)*.

The *value-added based simple labour productivity index* for production unit k , for period 1 relative to period 0, is defined as

$$ISLPROD^k_{VA}(1, 0) \equiv \frac{Q_{VA}^k(1, 0)}{L^{k1}/L^{k0}}, \quad (2.22)$$

which can then be interpreted as an index of real value added per unit of labour.

2.4.2 Levels

As one sees, some ‘level’-language has crept in. The bottom-up approach freely talks about productivity (change) in terms of levels. But what precisely are levels, and what is the relation between levels and indices? Intuitively, indices are just ratios of levels, so that it seems that the difference is merely in the kind of language one prefers. It appears, however, that a closer look is warranted.

For each production unit $k \in \mathcal{K}^t$ *real value added* is (ideally) defined as

$$RVA^k(t, b) \equiv VA^{kt}/P_{VA}^k(t, b); \quad (2.23)$$

that is, nominal value added at period t divided by (or, as one says, deflated by) a production-unit- k -specific value-added based price index for period t relative to a certain reference period b , where period b may or may not precede period 0. Notice that this definition tacitly assumes that production unit k , existing in period t , also existed or still exists in period b ; otherwise, deflation by a production-unit- k -specific index would be impossible. When production unit k does not exist in period b then for deflation a non-specific index must be used. On the complications thereby we will come back at a later stage.

The foregoing definition implies that

$$RVA^k(b, b) = VA^{kb}/P_{VA}^k(b, b) = VA^{kb}, \quad (2.24)$$

since any price index, whatever its functional form, returns the outcome 1 for the reference period. Thus, at the reference period b , real value added equals nominal value added.

For example, one easily checks that when $P_{VA}^k(t, b)$ is a Paasche-type double deflator, then real value added RVA^{kt} is period t value added at prices of period b (recall Balk 2010, Appendix B). The rather intricate form at the left-hand side of expression (2.23) serves to make clear that unlike VA^{kt} , which is an observable monetary magnitude, $RVA^k(t, b)$ is *the outcome of a function*. Though the outcome is also monetary, its magnitude depends on the reference period and the deflator chosen.

The first kind of dependence becomes clear by considering $RVA^k(t, b')$ for some $b' \neq b$. One immediately checks that $RVA^k(t, b')/RVA^k(t, b) = P_{VA}^k(t, b)/P_{VA}^k(t, b')$, which is a measure of the (k -specific value-added based) price difference between periods b' and b . Put otherwise, real value added depends critically on the price level of the reference period, which is the period for which nominal and real value added coincide.

As to the other dependence, it of course matters whether $P_{VA}^k(t, b)$ is a Paasche-type or a Laspeyres-type or a Fisher-type double deflator. Here the difference in general increases with increasing the time span between the periods t and b .

Another way of looking at real value added is to realize that, by using expression (2.17), $RVA^k(t, b) = VA^{kb}Q_{VA}^k(t, b)$. Put otherwise, real value added is a (normalized) quantity index.

Like real value added, *real primary, or capital-and-labour, input*, relative to reference period b , is (ideally) defined as deflated primary input cost,

$$X_{KL}^k(t, b) \equiv C_{KL}^{kt}/P_{KL}^k(t, b); \quad (2.25)$$

real capital input, relative to reference period b , is (ideally) defined as deflated capital cost,

$$X_K^k(t, b) \equiv C_K^{kt}/P_K^k(t, b); \quad (2.26)$$

and *real labour input*, relative to reference period b , is (ideally) defined as deflated labour cost,

$$X_L^k(t, b) \equiv C_L^{kt}/P_L^k(t, b), \quad (2.27)$$

Of course, similar observations as above apply to these two definitions. In particular, it is important to note that at the reference period b real primary input equals nominal input cost, $X_{KL}^k(b, b) = C_{KL}^{kb}$, real capital input equals nominal capital cost, $X_K^k(b, b) = C_K^{kb}$, and real labour input equals nominal labour cost, $X_L^k(b, b) = C_L^{kb}$.

It is important to observe that, whereas nominal values are additive, real values are generally not; that is, $X_{KL}^k(t, b) \neq X_K^k(t, b) + X_L^k(t, b)$ for $t \neq b$. It is easy to see, by combining expressions (2.25), (2.26) and (2.27), that requiring additivity means that the overall price index $P_{KL}^k(t, b)$ must be a second-stage Paasche index of the two subindices $P_K^k(t, b)$ and $P_L^k(t, b)$. When we are dealing with chained indices it is impossible to satisfy this requirement. An operationally feasible solution was proposed by Balk and Reich (2008).⁵

Using the foregoing building blocks, the *value-added based total factor productivity level* of production unit k at period t is defined as real value added divided by real primary input,

$$TFPROD_{VA}^k(t, b) \equiv \frac{RVA^k(t, b)}{X_{KL}^k(t, b)}. \quad (2.28)$$

Notice that numerator as well as denominator are expressed in the same price level, namely that of period b . Thus $TFPROD_{VA}^k(t, b)$ is a dimensionless variable.

The foregoing definition immediately implies that at the reference period b value-added based total factor productivity equals nominal value added divided by nominal primary input cost, $TFPROD_{VA}^k(b, b) = VA^{kb}/C_{KL}^{kb}$. Now recall the KL-VA

⁵Of course, a trivial solution would be to use the same deflator for all the nominal values. Such a strategy was proposed for the National Accounts by Durand (2004).

accounting identity (2.2) and assume that profit Π^{kt} is constrained to equal 0 for all production units at all time periods. Then reference period total factor productivity of all production units equals 1, $TFPROD_{VA}^k(b, b) = 1$ ($k \in \mathcal{K}^t$).

Likewise, the *value-added based labour productivity level* of unit k at period t is defined as real value added divided by real labour input,

$$LPROD_{VA}^k(t, b) \equiv \frac{RVA^k(t, b)}{X_L^k(t, b)}. \quad (2.29)$$

This is also a dimensionless variable. For the reference period b we obtain

$$LPROD_{VA}^k(b, b) = \frac{VA^{kb}}{C_L^{kb}} = \frac{VA^{kb}}{C_{KL}^{kb}} \frac{C_{KL}^{kb}}{C_L^{kb}}. \quad (2.30)$$

Hence, when profit $\Pi^{kt} = 0$ for all production units at all time periods then production unit k 's labour productivity at reference period b , $LPROD_{VA}^k(b, b)$ equals C_{KL}^{kb}/C_L^{kb} . This is the reciprocal of k 's labour cost share at period b .

In case the simple sum quantity index is used for labour, one obtains

$$LPROD_{VA}^k(t, b) = \frac{RVA^k(t, b)}{C_L^{kt}/P_L^k(t, b)} = \frac{RVA^k(t, b)}{C_L^{kb}Q_L^k(t, b)} = \frac{RVA^k(t, b)}{(C_L^{kb}/L^{kb})L^{kt}}, \quad (2.31)$$

where subsequently expressions (2.27) and (2.16) were used. The constant in the denominator, C_L^{kb}/L^{kb} , is the mean price of a unit of labour at reference period b .

The *simple value-added based labour productivity level* of unit k at period t is defined by

$$SLPROD_{VA}^k(t, b) \equiv \frac{RVA^k(t, b)}{L^{kt}}. \quad (2.32)$$

It is not unimportant to notice that its dimension is money-of-period- b per unit of labour.

2.4.3 Linking Levels and Indices

We now turn to the relation between levels and indices. One expects that taking the ratio of two levels would deliver an index, but let us have a look. Dividing unit k 's total factor or labour productivity level at period 1 by the same at period 0 delivers, using the various definitions and relations (2.17), (2.12) and (2.16),

$$\frac{TFPROD_{VA}^k(1, b)}{TFPROD_{VA}^k(0, b)} = \frac{Q_{VA}^k(1, b)/Q_{VA}^k(0, b)}{Q_{KL}^k(1, b)/Q_{KL}^k(0, b)}, \quad (2.33)$$

$$\frac{LPROD_{VA}^k(1, b)}{LPROD_{VA}^k(0, b)} = \frac{Q_{VA}^k(1, b)/Q_{VA}^k(0, b)}{Q_L^k(1, b)/Q_L^k(0, b)}, \quad (2.34)$$

$$\frac{SLPROD_{VA}^k(1, b)}{SLPROD_{VA}^k(0, b)} = \frac{Q_{VA}^k(1, b)/Q_{VA}^k(0, b)}{L^{k1}/L^{k0}}, \quad (2.35)$$

respectively. Surely, if $Q_{VA}^k(t, t')$, $Q_{KL}^k(t, t')$ and $Q_L^k(t, t')$ are well-behaving functions then the right-hand sides of expressions (2.33), (2.34) and (2.35) have the form of an output quantity index divided by an input quantity index, both for period 1 relative to period 0. When $b = 0, 1$ one easily checks that (2.33) reduces to $ITFPROD_{VA}^k(1, 0)$, that (2.34) reduces to $ILPROD_{VA}^k(1, 0)$, and that (2.35) reduces to $ISLPROD_{VA}^k(1, 0)$. But, when $b \neq 0, 1$, then

$$TFPROD_{VA}^k(1, b)/TFPROD_{VA}^k(0, b) = ITFPROD_{VA}^k(1, 0)$$

if and only if the quantity indices $Q_{VA}^k(t, t')$ and $Q_{KL}^k(t, t')$ are transitive (that is, satisfy the Circularity Test). Similarly,

$$LPROD_{VA}^k(1, b)/LPROD_{VA}^k(0, b) = ILPROD_{VA}^k(1, 0)$$

if and only if the quantity indices $Q_{VA}^k(t, t')$ and $Q_L^k(t, t')$ are transitive, and

$$SLPROD_{VA}^k(1, b)/SLPROD_{VA}^k(0, b) = ISLPROD_{VA}^k(1, 0)$$

if and only if the quantity index $Q_{VA}^k(t, t')$ is transitive.

Unfortunately, transitive quantity indices are in practice seldom used. Moreover, they would lead to price indices which fail some basic axioms.

2.4.4 When Not All the Data Are Accessible

The word ‘ideally’ was deliberately inserted in front of definitions (2.23), (2.25), (2.26) and (2.27). This word reflects the assumption that all the detailed price and quantity data, necessary to compile production-unit-specific price and quantity index numbers, are accessible. In practice, especially in the case of microdata, though the data are *available* at the enterprises—because revenue and cost are sums of quantities produced or used at certain unit prices—they are usually not *accessible* for researchers, due to the excessive cost of obtaining such data, their confidentiality, the response burden experienced by enterprises, or other reasons. In such cases researchers have to fall back at indices which are estimated for a higher aggregation level. This in turn means that real values are contaminated by differential price developments between the production units considered and the higher level aggregate.

In the extended version of this paper a number of situations are reviewed. In sectoral studies it appears that the way value added is deflated influences the distributions of the ensuing productivity levels; and the same holds at the input side.

A pervasive feature of microdata studies is the use of higher-level instead of production-unit specific deflators. There is some literature on the effect of using industry-level deflators instead of enterprise-level deflators on the estimation of production functions and the analysis of productivity change. See the early study of Abbott (1991) and, more recently, Mairesse and Jaumandreu (2005) and Foster et al. (2008). Of course, for such studies one needs enterprise-level price data, which severely limits the possibilities. In the literature, productivity based on revenue or value added deflated by an industry-level price index is sometimes called ‘revenue productivity’, to distinguish it from our concept that is then called ‘(physical) output productivity’.⁶

In a recent contribution Smeets and Warzynski (2013) found that physical productivity exhibited more dispersion than revenue productivity. A similar feature was unveiled by Eslava et al. (2013). In the last study it was also found that the correlation coefficient of the two measures was low. On the failure of revenue productivity measures to identify within-plant efficiency gains from exporting, see Marin and Voigtländer (2013). From the cross-sectional perspective this issue was studied by van Bieseboeck (2009).

2.5 Decompositions: Arithmetic Approach

Let us now assume that productivity levels, real output divided by real input, are somehow available.⁷ We denote the productivity level of unit k at period t by $PROD^{kt}$. Each production unit comes with some measure of relative size (importance) in the form of a weight θ^{kt} . These weights add up to 1 for each period, that is

$$\sum_{k \in \mathcal{K}^0} \theta^{k0} = \sum_{k \in \mathcal{K}^1} \theta^{k1} = 1. \quad (2.36)$$

⁶The distinction between revenue productivity and physical productivity is a central issue in the microdata study of Hsieh and Klenow (2009), where Indian, Chinese, and U.S. manufacturing plants/firms were compared over the period 1977–2005. However, they didn’t have access to plant/firm-level deflators. Using some theoretical reasoning, real value added was estimated as $RVA^k(t, b) = (VA^{kt})^{3/2}$, so that the ratio of physical productivity, calculated as $RVA^k(t, b)/X_{KL}^k(t, b)$, and revenue productivity, calculated as $VA^{kt}/X_{KL}^k(t, b)$, becomes equal to $(VA^{kt})^{1/2}$ ($k \in \mathcal{K}^t$). It comes as no surprise then that physical productivity exhibits more dispersion than revenue productivity.

⁷This section updates Balk (2003, Sect. 6).

We concentrate here on the productivity levels as introduced in the previous section; that is, $PROD^{kt}$ has the form of real value added divided by real primary input or real labour input. Then, ideally, the relative size measure θ^{kt} must be consistent with either of those measures. Though rather vague, this assumption is for the time being sufficient; we will return to this issue in Sect. 2.9.

The aggregate (or mean) productivity level at period t is quite naturally defined as the weighted arithmetic average of the unit-specific productivity levels, that is $PROD^t \equiv \sum_{k \in K^t} \theta^{kt} PROD^{kt}$, where the summation is taken over all production units existing at period t . The weighted geometric average, which is a natural alternative, as well as the weighted harmonic average, will be discussed in the next section.

Aggregate productivity change between periods 0 and 1 is then given by

$$PROD^1 - PROD^0 = \sum_{k \in K^1} \theta^{k1} PROD^{k1} - \sum_{k \in K^0} \theta^{k0} PROD^{k0}. \quad (2.37)$$

Given the distinction between continuing, exiting, and entering production units, as defined by expressions (2.10) and (2.11), expression (2.37) can be decomposed as

$$\begin{aligned} PROD^1 - PROD^0 &= \\ &\sum_{k \in N^1} \theta^{k1} PROD^{k1} \\ &+ \sum_{k \in C^{01}} \theta^{k1} PROD^{k1} - \sum_{k \in C^{01}} \theta^{k0} PROD^{k0} \\ &- \sum_{k \in X^0} \theta^{k0} PROD^{k0}. \end{aligned} \quad (2.38)$$

The first term at the right-hand side of the equality sign shows the contribution of entering units, the second and third term together show the contribution of continuing units, whereas the last term shows the contribution of exiting units. The contribution of continuing units, $\sum_{k \in C^{01}} \theta^{k1} PROD^{k1} - \sum_{k \in C^{01}} \theta^{k0} PROD^{k0}$, is the joint outcome of intra-unit productivity change, $PROD^{k1} - PROD^{k0}$, and inter-unit relative size change, $\theta^{k1} - \theta^{k0}$, for all $k \in C^{01}$. The problem of decomposing this joint outcome into the contributions of the two factors happens to be structurally similar to the index number (or indicator) problem. Whereas in index number theory we talk about prices, quantities, and commodities, we are here talking about sizes, productivity levels, and (continuing) production units.

It can thus be expected that in reviewing the various decomposition methods familiar names from index number theory, such as Laspeyres, Paasche, and Bennet, will turn up (see Balk 2008 for the nomenclature).

2.5.1 The First Three Methods

The first method decomposes the contribution of the continuing units into a Laspeyres-type contribution of intra-unit productivity change and a Paasche-type contribution of relative size change:

$$\begin{aligned}
 PROD^1 - PROD^0 = & \\
 \sum_{k \in \mathcal{N}^1} \theta^{k1} PROD^{k1} & \\
 + \sum_{k \in \mathcal{C}^{01}} \theta^{k0} (PROD^{k1} - PROD^{k0}) + \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) PROD^{k1} & \\
 - \sum_{k \in \mathcal{X}^0} \theta^{k0} PROD^{k0}. &
 \end{aligned} \tag{2.39}$$

The second term at the right-hand side of the equality sign relates to intra-unit productivity change and uses base period weights. It is therefore, using the language of index number theory, called a Laspeyres-type measure. The third term relates to relative size change and is weighted by comparison period productivity levels. It is therefore called a Paasche-type measure. This decomposition was used in the early microdata study of Baily, Hulten and Campbell (BHC) (1992).

One feature is important to notice. Disregard for a moment entering and exiting production units. Then aggregate productivity change is entirely due to continuing units, and is the sum of two terms. Suppose that all the units experience productivity increase, that is, $PROD^{k1} > PROD^{k0}$ for all $k \in \mathcal{C}^{01}$. Then aggregate productivity change is not necessarily positive, because the relative-size-change term $\sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) PROD^{k1}$ can exert a negative influence. This ‘paradox’ was extensively discussed by Fox (2012) and we will return to this issue in a later section.

Since base period and comparison period weights add up to 1, we can insert an arbitrary scalar a , and obtain

$$\begin{aligned}
 PROD^1 - PROD^0 = & \\
 \sum_{k \in \mathcal{N}^1} \theta^{k1} (PROD^{k1} - a) & \\
 + \sum_{k \in \mathcal{C}^{01}} \theta^{k0} (PROD^{k1} - PROD^{k0}) + \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) (PROD^{k1} - a) & \\
 - \sum_{k \in \mathcal{X}^0} \theta^{k0} (PROD^{k0} - a). &
 \end{aligned} \tag{2.40}$$

At this point it is useful to introduce some additional notation. Let $PROD^{\mathcal{X}^0} \equiv \sum_{k \in \mathcal{X}^0} \theta^{k0} PROD^{k0} / \sum_{k \in \mathcal{X}^0} \theta^{k0}$ be the mean productivity level of the exiting units, and let $PROD^{\mathcal{N}^1} \equiv \sum_{k \in \mathcal{N}^1} \theta^{k1} PROD^{k1} / \sum_{k \in \mathcal{N}^1} \theta^{k1}$ be the mean productivity level of the entering units. Then expression (2.40) can be written as

$$\begin{aligned}
PROD^1 - PROD^0 = & \\
& \left(\sum_{k \in \mathcal{N}^1} \theta^{k1} \right) (PROD^{\mathcal{N}^1} - a) \\
& + \sum_{k \in \mathcal{C}^{01}} \theta^{k0} (PROD^{k1} - PROD^{k0}) + \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) (PROD^{k1} - a) \\
& - \left(\sum_{k \in \mathcal{X}^0} \theta^{k0} \right) (PROD^{\mathcal{X}^0} - a).
\end{aligned} \tag{2.41}$$

Thus, entering units contribute positively to aggregate productivity change when their mean productivity level exceeds a , and exiting units contribute positively when their mean productivity level falls short of a . The net effect of entrance and exit is given by the sum of the first and the fourth right-hand side term, $(\sum_{k \in \mathcal{N}^1} \theta^{k1}) (PROD^{\mathcal{N}^1} - a) - (\sum_{k \in \mathcal{X}^0} \theta^{k0}) (PROD^{\mathcal{X}^0} - a)$. It is interesting to notice that this effect not only depends on relative importances and mean productivities, but also on the value chosen for the arbitrary scalar a . However, as we will see, there are a number of reasonable options here.

The second method uses a Paasche-type measure for intra-unit productivity change and a Laspeyres-type measure for relative size change. This leads to

$$\begin{aligned}
PROD^1 - PROD^0 = & \\
& \left(\sum_{k \in \mathcal{N}^1} \theta^{k1} \right) (PROD^{\mathcal{N}^1} - a) \\
& + \sum_{k \in \mathcal{C}^{01}} \theta^{k1} (PROD^{k1} - PROD^{k0}) + \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) (PROD^{k0} - a) \\
& - \left(\sum_{k \in \mathcal{X}^0} \theta^{k0} \right) (PROD^{\mathcal{X}^0} - a).
\end{aligned} \tag{2.42}$$

I am not aware of any application of this decomposition.

It is possible to avoid the choice between the Laspeyres-Paasche-type and the Paasche-Laspeyres-type decomposition. The third method uses for the contribution of both intra-unit productivity change and relative size change Laspeyres-type measures. However, this simplicity is counterbalanced by the necessity to introduce a covariance-type term:

$$\begin{aligned}
PROD^1 - PROD^0 = & \\
& \left(\sum_{k \in \mathcal{N}^1} \theta^{k1} \right) (PROD^{\mathcal{N}^1} - a) \\
& + \sum_{k \in \mathcal{C}^{01}} \theta^{k0} (PROD^{k1} - PROD^{k0}) + \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) (PROD^{k0} - a)
\end{aligned}$$

$$\begin{aligned}
& + \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0})(PROD^{k1} - PROD^{k0}) \\
& - \left(\sum_{k \in \mathcal{X}^0} \theta^{k0} \right) (PROD^{\mathcal{X}^0} - a).
\end{aligned} \tag{2.43}$$

In view of the overall Laspeyres-type perspective, a natural choice for the arbitrary scalar now seems to be $a = PROD^0$, the base period aggregate productivity level. This leads to the decomposition originally proposed by Haltiwanger (1997) and preferred by Foster, Haltiwanger and Krizan (FHK) (2001) (there called method 1). This method has been employed *inter alia* by Foster et al. (2006), Foster et al. (2008), and Collard-Wexler and de Loecker (2013).⁸ The FHK method will also be used in OECD's MultiProd project (OECD 2014).

Baldwin and Gu (2006) suggested to set $a = PROD^{\mathcal{X}^0}$, the base period mean productivity level of the exiting units. It is clear that then the final right-hand side term in expression (2.43) vanishes, and that the net effect of entrance and exit becomes equal to $(\sum_{k \in \mathcal{N}^1} \theta^{k1})(PROD^{\mathcal{N}^1} - PROD^{\mathcal{X}^0})$. It is as if the entering units have replaced the exiting units, and that the mean productivity surplus is all that matters.

Choosing $a = 0$ brings us back to the BHC decomposition. Nishida et al. (2014) provided interesting comparisons of the BHC and FHK decompositions on Chilean, Colombian and Slovenian micro-level data.

2.5.2 *Interlude: The TRAD, CSLS, and GEA Decompositions*

Let us pause for a while at this expression and consider the case where there is neither exit nor entry; that is $\mathcal{K}^0 = \mathcal{K}^1 = \mathcal{C}^{01}$. Then expression (2.43) reduces to

$$\begin{aligned}
PROD^1 - PROD^0 &= \\
& \sum_{k \in \mathcal{C}^{01}} \theta^{k0} (PROD^{k1} - PROD^{k0}) \\
& + \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) (PROD^{k0} - a) \\
& + \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) (PROD^{k1} - PROD^{k0}).
\end{aligned} \tag{2.44}$$

⁸ Altomonte and Nicolini (2012) applied the FHK method to aggregate price-cost margin change. For any individual production unit the price-cost margin was defined as nominal cash flow (= value added minus labour cost) divided by nominal revenue, CF^{kt}/R^{kt} . These margins were weighted by market shares $R^{kt}/\sum_{k \in \mathcal{K}'} R^{kt}$ ($k \in \mathcal{K}'$).

In order to transform to (forward looking) percentage changes (aka growth rates) both sides of this expression are divided by $PROD^0$, which delivers

$$\begin{aligned} \frac{PROD^1 - PROD^0}{PROD^0} &= \\ \sum_{k \in C^{01}} \theta^{k0} \frac{PROD^{k0}}{PROD^0} &\left(\frac{PROD^{k1} - PROD^{k0}}{PROD^{k0}} \right) \\ + \sum_{k \in C^{01}} (\theta^{k1} - \theta^{k0}) \frac{PROD^{k0} - a}{PROD^0} \\ + \sum_{k \in C^{01}} \frac{PROD^{k0}}{PROD^0} (\theta^{k1} - \theta^{k0}) \left(\frac{PROD^{k1} - PROD^{k0}}{PROD^{k0}} \right). \end{aligned} \quad (2.45)$$

Now consider *simple* labour productivity, that is, real value added per unit of labour; thus $PROD^{kt} = SLPROD_{VA}^k(t, b) \equiv RVA^k(t, b)/L^{kt}$ ($k \in C^{01}$). Let the relative size of a production unit be given by its labour share; that is, $\theta^{kt} \equiv \frac{L^{kt}}{\sum_{k \in C^{01}} L^{kt}}$ ($k \in C^{01}$). It is straightforward to check that then the weights occurring in the first right-hand side term expression (2.45), $\theta^{k0} \frac{PROD^{k0}}{PROD^0}$, reduce to real-value-added shares, $\frac{RVA^k(0, b)}{\sum_{k \in C^{01}} RVA^k(0, b)}$ ($k \in C^{01}$), so that

$$\begin{aligned} \frac{PROD^1 - PROD^0}{PROD^0} &= \\ \sum_{k \in C^{01}} \frac{RVA^k(0, b)}{\sum_{k \in C^{01}} RVA^k(0, b)} &\left(\frac{PROD^{k1} - PROD^{k0}}{PROD^{k0}} \right) \\ + \sum_{k \in C^{01}} (\theta^{k1} - \theta^{k0}) \frac{PROD^{k0} - a}{PROD^0} \\ + \sum_{k \in C^{01}} \frac{PROD^{k0}}{PROD^0} (\theta^{k1} - \theta^{k0}) \left(\frac{PROD^{k1} - PROD^{k0}}{PROD^{k0}} \right). \end{aligned} \quad (2.46)$$

In view of the fact that $\sum_{k \in C^{01}} (\theta^{k1} - \theta^{k0}) = 0$, expression (2.46) can also be written as

$$\begin{aligned} \frac{PROD^1 - PROD^0}{PROD^0} &= \\ \sum_{k \in C^{01}} \frac{RVA^k(0, b)}{\sum_{k \in C^{01}} RVA^k(0, b)} &\left(\frac{PROD^{k1} - PROD^{k0}}{PROD^{k0}} \right) \end{aligned}$$

$$\begin{aligned}
& + \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) \frac{PROD^{k0} - a}{PROD^0} \\
& + \sum_{k \in \mathcal{C}^{01}} \frac{PROD^{k0}}{PROD^0} (\theta^{k1} - \theta^{k0}) \left(\frac{PROD^{k1} - PROD^{k0} - a'}{PROD^{k0}} \right), \quad (2.47)
\end{aligned}$$

for another arbitrary scalar a' . Now, choosing $a = 0$ and $a' = 0$ yields the TRAD(itional) way of decomposing aggregate labour productivity change into contributions of the various industries, according to three main sources: a within-sector effect, a reallocation level effect, and a reallocation growth effect respectively (see for various other names and their provenances De Avillez 2012). Choosing $a = PROD^0$ and $a' = PROD^1 - PROD^0$ yields the CSLS decomposition (which has been developed at the Centre for the Study of Living Standards).

Finally, let the relative size of a production unit be given by its combined labour and relative price share; that is, $\theta^{kt} \equiv \frac{L^{kt}}{\sum_{k \in \mathcal{C}^{01}} L^{k\bar{t}}} \frac{P_{VA}^k(t, b)}{P_{VA}^K(t, b)}$ ($k \in \mathcal{C}^{01}$), where $P_{VA}^K(t, b)$ is some non- k -specific deflator. Notice that these weights do not add up to 1. It is straightforward to check that in this case the weights occurring in the first right-hand side term expression (2.45), $\theta^{k0} \frac{PROD^{k0}}{PROD^0}$, reduce to nominal-value-added shares, $\frac{VA^{k0}}{\sum_{k \in \mathcal{C}^{01}} VA^{k0}}$ ($k \in \mathcal{C}^{01}$), so that

$$\begin{aligned}
& \frac{PROD^1 - PROD^0}{PROD^0} = \\
& \sum_{k \in \mathcal{C}^{01}} \frac{VA^{k0}}{\sum_{k \in \mathcal{C}^{01}} VA^{k0}} \left(\frac{PROD^{k1} - PROD^{k0}}{PROD^{k0}} \right) \\
& + \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) \frac{PROD^{k0} - a}{PROD^0} \\
& + \sum_{k \in \mathcal{C}^{01}} \frac{PROD^{k0}}{PROD^0} (\theta^{k1} - \theta^{k0}) \left(\frac{PROD^{k1} - PROD^{k0} - a'}{PROD^{k0}} \right), \quad (2.48)
\end{aligned}$$

where a and a' are arbitrary scalars. For $a = 0$ and $a' = 0$ this appears to be the Generalized Exactly Additive Decomposition (GEAD), going back to Tang and Wang (2004) and explored by Dumagan (2013).

De Avillez (2012) provided an interesting empirical comparison of TRAD, CSLS, and GEAD. He found that “despite some similarities, all three decomposition formulas paint very different pictures of which sectors drove labour productivity growth in the Canadian business sector during the 2000–2010 period.” The difference between TRAD and CSLS not unexpectedly hinges on the role played by the scalars a and a' . Varying a and/or a' implies varying magnitudes of the two reallocation effects, not at the aggregate level—because the sums are invariant—but at the level of individual production units (in his case, industries).

The difference between TRAD and CSLS on the one hand and GEAD on the other evidently hinges on the absence or presence of relative price levels in the sectoral measures of importance.⁹ De Avillez found it “impossible to say which set of estimates provides a more accurate picture of economic reality because the GEAD formula is, ultimately, measuring something very different from the TRAD and CSLS formulas.” I concur insofar this conclusion only means that the answer cannot be found within the bottom-up perspective. The top-down perspective is required to obtain a decision.

2.5.3 *The Fourth and Fifth Method*

Let us now return to expression (2.43). Instead of the Laspeyres perspective, one might as well use the Paasche perspective. The covariance-type term accordingly appears with a negative sign. Thus, the fourth decomposition is

$$\begin{aligned}
 PROD^1 - PROD^0 = & \\
 & \left(\sum_{k \in \mathcal{N}^1} \theta^{k1} \right) (PROD^{\mathcal{N}^1} - a) \\
 & + \sum_{k \in \mathcal{C}^{01}} \theta^{k1} (PROD^{k1} - PROD^{k0}) + \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) (PROD^{k1} - a) \\
 & - \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) (PROD^{k1} - PROD^{k0}) \\
 & - \left(\sum_{k \in \mathcal{X}^0} \theta^{k0} \right) (PROD^{\mathcal{X}^0} - a).
 \end{aligned} \tag{2.49}$$

The natural choice for a would now be $PROD^1$, the comparison period aggregate productivity level. Choosing $a = PROD^{\mathcal{N}^1}$ would lead to disappearance of the entry effect. The net effect of entrance and exit then becomes equal to $(\sum_{k \in \mathcal{X}^0} \theta^{k0})(PROD^{\mathcal{N}^1} - PROD^{\mathcal{X}^0})$. It is left to the reader to explore the analogs to expressions (2.47) and (2.48) by using backward looking percentage changes. I am not aware of any empirical application of this decomposition.

The fifth method avoids the Laspeyres-Paasche dichotomy altogether, by using the symmetric Bennet-type method. This amounts to taking the arithmetic average of the first and the second method. The covariance-type term then disappears. Thus,

$$PROD^1 - PROD^0 =$$

⁹See also Reinsdorf (2015). He considered a convex combination of CSLS with price reference periods $b = 0$ and $b = 1$.

$$\begin{aligned}
& \left(\sum_{k \in \mathcal{N}^1} \theta^{k1} \right) (PROD^{\mathcal{N}^1} - a) \\
& + \sum_{k \in \mathcal{C}^{01}} \frac{\theta^{k0} + \theta^{k1}}{2} (PROD^{k1} - PROD^{k0}) \\
& + \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) \left(\frac{PROD^{k0} + PROD^{k1}}{2} - a \right) \\
& - \left(\sum_{k \in \mathcal{X}^0} \theta^{k0} \right) (PROD^{\mathcal{X}^0} - a).
\end{aligned} \tag{2.50}$$

With respect to the scalar a there are several options available in the literature. A rather natural choice is $a = (PROD^0 + PROD^1)/2$, the overall two-period mean aggregate productivity level. Then, entering units contribute positively to aggregate productivity change if their mean productivity level is above this overall mean. Exiting units contribute positively if their mean productivity level is below the overall mean. Continuing units can contribute positively in two ways: if their productivity level increases, or if the units with productivity levels above (below) the overall mean increase (decrease) in relative size. This decomposition basically corresponds to the one used in the early microdata study of Griliches and Regev (GR) (1995). Because of its symmetry it is widely preferred. Moreover, Foster et al. (2001) argue that the GR method (there called method 2) is presumably less sensitive to (random) measurement errors than the asymmetric FHK method. The GR method was employed by Baily et al. (2001) and Foster et al. (2008).

But other choices are also plausible. Baldwin and Gu (BG) (2006) suggested to set $a = PROD^{\mathcal{X}^0}$, the base period mean productivity level of the exiting production units. Then, as we have seen before, the last term of expression (2.50) disappears. Put otherwise, entering units are seen as displacing exiting units, contributing positively to aggregate productivity change insofar their mean productivity level exceeds that of the exiting units.

Baldwin and Gu (2008) considered two alternatives, to be applied to different types of industries. The first is to set a equal to the base period mean productivity level of the continuing units that are contracting; that is, the units $k \in \mathcal{C}^{01}$ for which $\theta^{k1} < \theta^{k0}$. The second is to set a equal to the base period mean productivity level of the continuing units that are expanding.

Balk and Hoogenboom-Spijker (2003) compared the five methods, defined by expressions (2.41), (2.42), (2.43), (2.49), and (2.50) respectively, on micro-level data of the Netherlands manufacturing industry over the period 1984–1999. Though there appeared to be appreciable differences between the various decompositions, the pervasive fact was the preponderance of the productivity change of the continuing units (or, the ‘within’ term).

2.5.4 Another Five Methods

A common feature of the five decomposition methods discussed hitherto is that the productivity levels of exiting and entering production units are compared to a single overall benchmark level a , for which a number of options is available. It seems more natural to compare the productivity levels of exiting units to the mean level of the continuing units at the base period—which is the period of exit, and to compare the productivity levels of entering units to the mean level of the continuing units at the comparison period—which is the period of entrance.

Thus, let the aggregate productivity level of the continuing production units at period t be defined as $PROD^{C^{01}t} \equiv \sum_{k \in C^{01}} \theta^{kt} PROD^{kt} / \sum_{k \in C^{01}} \theta^{kt}$ ($t = 0, 1$). Since the weights θ^{kt} add up to 1 for both periods—see expression (2.36)—expression (2.37) can be decomposed as

$$\begin{aligned} PROD^1 - PROD^0 = & \\ & \left(\sum_{k \in N^1} \theta^{k1} \right) (PROD^{N^1} - PROD^{C^{01}1}) \\ & + PROD^{C^{01}1} - PROD^{C^{01}0} \\ & - \left(\sum_{k \in X^0} \theta^{k0} \right) (PROD^{X^0} - PROD^{C^{01}0}). \end{aligned} \quad (2.51)$$

This expression tells us that entering units contribute positively to aggregate productivity change if their mean productivity level is above that of the continuing units at the entrance period. Similarly, exiting units contribute positively if their mean productivity level is below that of the continuing units at the period of exit.

Let the relative size of continuing units be defined by $\tilde{\theta}^{kt} \equiv \theta^{kt} / \sum_{k \in C^{01}} \theta^{kt}$ ($k \in C^{01}; t = 0, 1$). The contribution of the continuing units to aggregate productivity change can then be written as

$$PROD^{C^{01}1} - PROD^{C^{01}0} = \sum_{k \in C^{01}} \tilde{\theta}^{k1} PROD^{k1} - \sum_{k \in C^{01}} \tilde{\theta}^{k0} PROD^{k0}, \quad (2.52)$$

which has the same structure as the second and third term of expression (2.38), the difference being that the weights now add up to 1; that is, $\sum_{k \in C^{01}} \tilde{\theta}^{kt} = 1$ ($t = 0, 1$). Thus the five methods discussed earlier can simply be repeated on the right-hand side of expression (2.51). The first four, asymmetric, methods are left to the reader. The symmetric Bennet decomposition delivers the following result,

$$\begin{aligned} PROD^1 - PROD^0 = & \\ & \left(\sum_{k \in N^1} \theta^{k1} \right) (PROD^{N^1} - PROD^{C^{01}1}) \\ & + \sum_{k \in C^{01}} \frac{\tilde{\theta}^{k0} + \tilde{\theta}^{k1}}{2} (PROD^{k1} - PROD^{k0}) \end{aligned}$$

$$\begin{aligned}
& + \sum_{k \in \mathcal{C}^{01}} (\tilde{\theta}^{k1} - \tilde{\theta}^{k0}) \left(\frac{PROD^{k0} + PROD^{k1}}{2} - a \right) \\
& - \left(\sum_{k \in \mathcal{X}^0} \theta^{k0} \right) (PROD^{\mathcal{X}^0} - PROD^{\mathcal{C}^{01}0}). \tag{2.53}
\end{aligned}$$

The first right-hand side term of this expression refers to the entering production units. As we see, its magnitude is determined by the period 1 share of the entrants and the productivity gap with the continuing units. The last right-hand side term refers to the exiting production units. The magnitude of this term depends on the share of the exiting units and the productivity gap with the continuing units. The second and third term refer to the continuing production units. They may contribute positively in two ways: if their productivity levels on average increase, or if the units with mean productivity levels above (or below) the scalar a increase (or decrease) in relative size.

Notice that the third term is the only place where an arbitrary scalar a can be inserted, since the relative weights of the continuing production units add up to 1 in both periods. Though the term itself is invariant, the unit-specific components $(\tilde{\theta}^{k1} - \tilde{\theta}^{k0})((PROD^{k0} + PROD^{k1})/2 - a)$ are not.

This decomposition was developed by Diewert and Fox (DF) (2010), the discussion paper version of which was published in 2005. Though in a different context—the development of shares of labour types in plant employment—the DF decomposition in the field of productivity measurement can be detected in Vainiomäki (1999). Currently there are hardly any empirical applications known.¹⁰

If there are no exiting or entering units, that is, $\mathcal{K}^0 = \mathcal{K}^1 = \mathcal{C}^{01}$, then the DF method (2.53) as well as the GR method (2.50) reduce to the simple Bennet-type decomposition.

¹⁰Though Kirwan et al. (2012) contend to use the DF decomposition, it appears that their analysis is simply based on expression (2.37). The part relating to continuing units is replaced by a weighted sum of production function based unit-specific productivity changes plus residuals. Böckerman and Maliranta (2007) used the DF decomposition for the analysis of value-added based simple labour productivity and total factor productivity. Kauhanen and Maliranta (2012) applied a two-stage DF decomposition to mean wage change.

2.5.5 Provisional Evaluation

The overview provided in the foregoing subsections hopefully demonstrates a number of things, the first and most important of which is that there is no unique decomposition of aggregate productivity change as defined by expression (2.37).¹¹

Second, one should be careful with reifying the different components, in particular the covariance-type term, since this term can be considered as a mere artefact arising from the specific (Laspeyres- or Paasche-) perspective chosen.

Third, the undetermined character of the scalar a lends additional arbitrariness to the first set of five decompositions. At the aggregate level it is easily seen that letting a tend to 0 will lead to a larger contribution of the entering units, the exiting units, and the size change of continuing units, at the expense of intra-unit productivity change. The advantage of the second set of five decompositions, among which the symmetric DF method, is that the distribution of these four parts is kept unchanged. The remaining arbitrariness in expression (2.53) is in the size-change term and materializes only at the level of individual continuing production units.

Fourth, what counts as ‘entrant’ or ‘exiting unit’ depends not only on the length of the time span between the periods 0 and 1, but also on the length of the periods itself and the observation thresholds employed in sampling.

All in all it can be expected that the outcome of any decomposition exercise depends to some extent on the particular method favoured by the researcher. This is not a problem as long as he or she realizes this and let the favoured results be accompanied by some alternatives.

Finally, as demonstrated in the previous section, the productivity levels $PROD^{kt}$ depend on the price reference period of the deflators used. In particular this holds for the simple labour productivity levels $RVA^k(t, b)/L^{kt}$. This dependence obviously extends to aggregate productivity change $PROD^1 - PROD^0$. To mitigate its effect, one considers instead the forward-looking growth rate of aggregate productivity $(PROD^1 - PROD^0)/PROD^0$ and its decomposition, obtained by dividing each term by $PROD^0$. It would of course be equally justified to consider the backward-looking growth rate $(PROD^1 - PROD^0)/PROD^1$. A symmetric growth rate is obtained when the difference $PROD^1 - PROD^0$ is divided by a mean of $PROD^0$ and $PROD^1$. When the logarithmic mean¹² is used, one obtains

$$\frac{PROD^1 - PROD^0}{L(PROD^0, PROD^1)} = \ln(PROD^1 / PROD^0), \quad (2.54)$$

¹¹This non-uniqueness should not come as a surprise and finds its parallel in index number theory (see Balk 2008) and in so-called structural decomposition analysis (widely used in input-output analysis; see Dietzenbacher and Los 1998).

¹²The logarithmic mean, for any two strictly positive real numbers a and b , is defined by $L(a, b) \equiv (a - b) / \ln(a/b)$ if $a \neq b$ and $L(a, a) \equiv a$. See Balk (2008, pp. 134–136) for a discussion of its properties, one of these being that $(ab)^{1/2} \leq L(a, b) \leq (a + b)/2$.

which can be interpreted as a percentage change.¹³ However, its decomposition still contains differences such as $PROD^{k1} - PROD^{k0}$, which of course can be transformed into logarithmic differences but at the expense of getting pretty complicated weights.

Thus this calls for going geometric right from the start; which is the topic of the next section.

2.6 Decompositions: Geometric and Harmonic Approach

In the geometric approach the aggregate productivity level is defined as a weighted *geometric average* of the unit-specific productivity levels, that is $PROD^t \equiv \prod_{k \in K^t} (PROD^{kt})^{\theta^{kt}}$. This is equivalent to defining $\ln PROD^t \equiv \sum_{k \in K^t} \theta^{kt} \ln PROD^{kt}$, which implies that, by replacing $PROD$ by $\ln PROD$, the entire story of the previous section can be repeated.

The advantage of decomposing $\ln PROD^1 - \ln PROD^0$ over decomposing $PROD^1 - PROD^0$ is that a logarithmic change can be interpreted immediately as a percentage change. The disadvantage is that, as an aggregate level measure, a geometric mean $\prod_{k \in K^t} (PROD^{kt})^{\theta^{kt}}$ is less easy to understand than an arithmetic mean $\sum_{k \in K^t} \theta^{kt} PROD^{kt}$. We let the top-down approach here advise which mean should be preferred; see the extended version of this paper.

The Geometric DF decomposition was applied by Hytyinen and Maliranta (2013). They extended the decomposition to deal with age groups of firms. There are of course also geometric versions of the GR, FHK, and BG decompositions. Baldwin and Gu (2011) compared these on Canadian retail trade and manufacturing industry microdata over the period 1984–1998. As in the comparative study of Balk and Hoogenboom-Spijker (2003), they found that in manufacturing the ‘within’ term was dominant. However, in retail trade the net effect of entry and exit appeared more important.

In the harmonic approach the aggregate productivity level is defined as a weighted *harmonic average* of the unit-specific productivity levels, that is $PROD^t \equiv (\sum_{k \in K^t} \theta^{kt} (PROD^{kt})^{-1})^{-1}$. Though the literature does hardly pay any attention to this option, in the extended version of this paper it will be shown that there are situations in which this type of average rather naturally emerges. An example is provided by Böckerman and Maliranta (2012). Though these authors were primarily concerned with the evolution of the aggregate real labour share through time, it turns out that their analysis is equivalent to an Harmonic DF decomposition on aggregate labour productivity, defined as weighted harmonic mean of $LPROD_{VA}^k(t, t-1)$ with weights defined as real value added shares at period t .

¹³Since $\ln(a/a') = \ln(1 + (a - a')/a') \approx (a - a')/a'$ when $(a - a')/a'$ is small.

2.7 Monotonicity

As already preluded to, all the definitions of aggregate productivity change, whether arithmetic, geometric or harmonic, suffer from what Fox (2012) called the “monotonicity problem” or “paradox”.

Again, disregard for a moment entering and exiting production units. Then aggregate productivity change is entirely due to continuing units, and is the combination (sum, product, or harmonic sum, respectively) of two terms. Suppose that all the units experience productivity increase, that is, $PROD^{k1} > PROD^{k0}$ for all $k \in C^0$. Then the ‘within’ term in the DF decomposition (2.53) and in the Geometric DF decomposition is positive, and in the Harmonic DF decomposition negative. However, aggregate productivity change is not necessarily positive, because the relative-size-change terms, can exert a counterbalancing influence.

Fox (2012) noticed that the term $\sum_{k \in C^0} (\theta^{k0} + \theta^{k1})(PROD^{k1} - PROD^{k0})/2$ as such has the desired monotonicity property, and proposed to extend this measure to the set $C^0 \cup X^0 \cup N^1 = K^0 \cup N^1 = X^0 \cup K^1$. Aggregate productivity change is then defined as

$$\Delta PROD_{Fox}(1, 0) \equiv \sum_{k \in C^0 \cup X^0 \cup N^1} \frac{\theta^{k0} + \theta^{k1}}{2} (PROD^{k1} - PROD^{k0}). \quad (2.55)$$

Now, for all exiting production units, $k \in X^0$, it is evidently the case that in the later period 1 those units have size zero; that is, $\theta^{k1} = 0$. It is then rather natural to set their virtual productivity level also equal to zero; that is, $PROD^{k1} = 0$. Likewise, entering units, $k \in N^1$, have size zero in the earlier period 0; that is, $\theta^{k0} = 0$. Their virtual productivity level at that period is also set equal to zero; that is, $PROD^{k0} = 0$. Then expression (2.55) can be decomposed as

$$\begin{aligned} \Delta PROD_{Fox}(1, 0) &= \\ (1/2) \sum_{k \in N^1} \theta^{k1} PROD^{k1} &+ \sum_{k \in C^0} \frac{\theta^{k0} + \theta^{k1}}{2} (PROD^{k1} - PROD^{k0}) \\ &- (1/2) \sum_{k \in X^0} \theta^{k0} PROD^{k0}. \end{aligned} \quad (2.56)$$

Unfortunately, there is no geometric or harmonic analog to expressions (2.55) and (2.56), because the logarithm or reciprocal of a zero productivity level is infinity. By using logarithmic means, one obtains

$$\begin{aligned}
\Delta PROD_{Fox}(1, 0) = & \\
& (1/2) \sum_{k \in \mathcal{N}^1} \theta^{k1} PROD^{k1} \\
& + \sum_{k \in \mathcal{C}^{01}} \frac{\theta^{k0} + \theta^{k1}}{2} L(PROD^{k0}, PROD^{k1}) \ln \left(PROD^{k1} / PROD^{k0} \right) \\
& - (1/2) \sum_{k \in \mathcal{X}^0} \theta^{k0} PROD^{k0},
\end{aligned} \tag{2.57}$$

which, however, does not provide any advantage vis à vis expression (2.56).

It is interesting to compare Fox's proposal to the symmetric decomposition (2.50) with $\alpha = 0$. It turns out that

$$\begin{aligned}
PROD^1 - PROD^0 = & \\
& \Delta PROD_{Fox}(1, 0) \\
& + (1/2) \sum_{k \in \mathcal{N}^1} \theta^{k1} PROD^{k1} \\
& + \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) \frac{PROD^{k0} + PROD^{k1}}{2} \\
& - (1/2) \sum_{k \in \mathcal{X}^0} \theta^{k0} PROD^{k0}.
\end{aligned} \tag{2.58}$$

It is remarkable that of the entire contribution of entering and exiting production units to $PROD^1 - PROD^0$, half is considered as productivity change and half as non-productivity change. It is difficult to envisage a solid justification for this.

2.8 The Olley-Pakes Decomposition

Though aggregate, or weighted mean, productivity levels are interesting, researchers are also interested in the *distribution* of the unit-specific levels $PROD^{kt}$ ($k \in \mathcal{K}^t$), and the change of such distributions over time, a good example being Bartelsman and D'hrymes (1998). Given the relative size measures θ^{kt} —which are adding up to 1—a natural question is whether high or low productivity of a unit goes together with high or low size. Are big firms more productive than small firms? Or are the most productive firms to be found among the smallest? Questions multiply when the time dimension is taken into account. Does the ranking of a particular production unit in the productivity distribution sustain through time? Are firms ranked somewhere in a particular period likely to rank higher or lower in the next period? Is there a relation

with the age, however determined, of the production units? Do the productivity distributions, and the behaviour of the production units, differ over the industries?

When it comes to size a natural measure to consider is the covariance of weights and productivity levels. Let $\#(\mathcal{K}^t)$ be the number of units in \mathcal{K}^t , let $\overline{PROD}^t \equiv \sum_{k \in \mathcal{K}^t} PROD^{kt} / \#(\mathcal{K}^t)$ be the unweighted mean of the productivity levels, and let $\bar{\theta}^t \equiv \sum_{k \in \mathcal{K}^t} \theta^{kt} / \#(\mathcal{K}^t) = 1 / \#(\mathcal{K}^t)$ be the unweighted mean of the weights. One then easily checks that

$$\sum_{k \in \mathcal{K}^t} (\theta^{kt} - \bar{\theta}^t)(PROD^{kt} - \overline{PROD}^t) = PROD^t - \overline{PROD}^t. \quad (2.59)$$

This is a particular instance of a general relation derived by Bortkiewicz in 1923/1924. Bortkiewicz showed that the difference between two differently weighted means has the form of a covariance. Interesting applications can be found in index number theory (see Balk 2008).

Olley and Pakes (OP) (1996, 1290) rearranged this relation to the form

$$PROD^t = \overline{PROD}^t + \sum_{k \in \mathcal{K}^t} (\theta^{kt} - \bar{\theta}^t)(PROD^{kt} - \overline{PROD}^t) \quad (2.60)$$

and provided an interpretation which has been repeated, in various forms, by many researchers.¹⁴ The interpretation usually goes like this: There is some event (say, a certain technological innovation or some other shock) that gives rise to a productivity level \overline{PROD}^t ; but this productivity level is transformed into an aggregate level $PROD^t$ by means of a mechanism called *reallocation*, the extent of which is measured by the covariance term in expression (2.60). So it seems that the aggregate productivity level $PROD^t$ is ‘caused’ by two factors, a productivity shock and a reallocation.¹⁵

I propose to call this the Olley-Pakes *fallacy*, because there are not at all two factors. Expression (2.59) is a mathematical identity: reallocation, defined as a covariance, is identically equal to the difference of two means, a weighted and an unweighted one. All that expression (2.60) does is featuring the unweighted mean rather than the weighted mean as the baseline variable.

I don’t dispute the usefulness of studying time-series or cross-sections of covariances such as we see at the left-hand side of expression (2.59). Notice that by replacing $PROD$ by $\ln PROD$ or $1/PROD$ one obtains a geometric or harmonic variant respectively. Additional insight can be obtained when one replaces produc-

¹⁴It is straightforward to generalize the OP decomposition to the case where the ensemble \mathcal{K}^t consists of a number of disjunct groups. The right-hand side of expression (2.60) then becomes the sum of a between-groups covariance and for each group an unweighted mean productivity level and a within-group covariance. Collard-Wexler and de Loecker (2013) considered a case of two groups.

¹⁵In Foster et al. (2001)’s article the OP decomposition, expression (2.60), was called method 3.

tivity levels $PROD^{kt}$ by productivity changes, measured as differences $PROD^{k1} - PROD^{k0}$ or percentage changes $\ln(PROD^{k1}/PROD^{k0})$. As a descriptive device this is wonderful, especially for comparing ensembles (industries, economies)—see for instance Lin and Huang (2012) where such covariances are regressed on several background variables. In the cross-country study of Bartelsman et al. (2013) within-industry covariances between size and productivity play a key role.¹⁶

The OP decomposition, expression (2.60), can of course be used to decompose aggregate productivity change $PROD^1 - PROD^0$ into two terms, the first being $\overline{PROD}^1 - \overline{PROD}^0$, and the second being the difference of two covariance terms. But then we are unable to distinguish between the contributions of exiting, continuing, and entering production units. Thus, it is advisable to restrict the OP decomposition to the continuing units, and substitute into expression (2.51). Doing this results in the following expression,

$$\begin{aligned} PROD^1 - PROD^0 = & \\ & \left(\sum_{k \in \mathcal{N}^1} \theta^{k1} \right) (PROD^{\mathcal{N}^1} - PROD^{\mathcal{C}^{01}1}) \\ & + \overline{PROD}^{\mathcal{C}^{01}1} - \overline{PROD}^{\mathcal{C}^{01}0} \\ & + \sum_{k \in \mathcal{C}^{01}} (\tilde{\theta}^{k1} - 1/\#(\mathcal{C}^{01})) (PROD^{k1} - \overline{PROD}^{\mathcal{C}^{01}1}) \\ & - \sum_{k \in \mathcal{C}^{01}} (\tilde{\theta}^{k0} - 1/\#(\mathcal{C}^{01})) (PROD^{k0} - \overline{PROD}^{\mathcal{C}^{01}0}) \\ & - \left(\sum_{k \in \mathcal{X}^0} \theta^{k0} \right) (PROD^{\mathcal{X}^0} - PROD^{\mathcal{C}^{01}0}), \end{aligned} \quad (2.61)$$

where $PROD^{\mathcal{C}^{01}t}$ is the weighted mean productivity level and $\overline{PROD}^{\mathcal{C}^{01}t} \equiv \sum_{k \in \mathcal{C}^{01}} PROD^{kt}/\#(\mathcal{C}^{01})$ is the unweighted mean productivity level of the continuing units at period t ($t = 0, 1$); $\#(\mathcal{C}^{01})$ is the number of those units.

This then is the decomposition proposed by Melitz and Polanec (2015). Their paper contains an interesting empirical comparison of the GR method (2.50), the FHK method (2.43), and the extended OP method (2.61). Hansell and Nguyen (2012) compared the BG method (2.50), the DF method (2.53), and the extended OP method (2.61). Again, their overall conclusion on Australian data concerning the 2002–2010 period was that the “dominant source of labour productivity growth in manufacturing and professional services is from within firms.”

¹⁶See also the special issue on “Misallocation and Productivity” of the *Review of Economic Dynamics* 16(1)(2013). There appears to be no unequivocal definition of ‘misallocation’. In OECD (2014) at least three different concepts can be detected.

Wolf (2011, pp. 21–25), see also Bartelsman and Wolf (2014), used the OP decomposition to enhance the GR decomposition. By substituting expression (2.60) into expression (2.50), with $a = (PROD^0 + PROD^1)/2$, one obtains

$$\begin{aligned}
 PROD^1 - PROD^0 &= \\
 &\left(\sum_{k \in \mathcal{N}^1} \theta^{k1} \right) \left(PROD^{\mathcal{N}^1} - \frac{PROD^0 + PROD^1}{2} \right) \\
 &+ \sum_{k \in \mathcal{C}^{01}} \frac{\theta^{k0} + \theta^{k1}}{2} (PROD^{k1} - PROD^{k0}) \\
 &+ \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) \left(\frac{PROD^{k0} + PROD^{k1}}{2} - \frac{\overline{PROD}^0 + \overline{PROD}^1}{2} \right) \\
 &- \sum_{k \in \mathcal{C}^{01}} (\theta^{k1} - \theta^{k0}) \left(\sum_{k \in \mathcal{K}^0} (\theta^{k0} - \bar{\theta}^0) (PROD^{k0} - \overline{PROD}^0) \right. \\
 &\quad \left. + \sum_{k \in \mathcal{K}^1} (\theta^{k1} - \bar{\theta}^1) (PROD^{k1} - \overline{PROD}^1) \right) / 2 \\
 &- \left(\sum_{k \in \mathcal{X}^0} \theta^{k0} \right) \left(PROD^{\mathcal{X}^0} - \frac{PROD^0 + PROD^1}{2} \right). \tag{2.62}
 \end{aligned}$$

As one sees, the original GR ‘between’ term, the third right-hand side term in expression (2.50), is split into two parts. The first part, which is the third right-hand side term in the last expression, is relatively easy to understand: it is still a covariance between size changes and mean productivity levels. The second part, which is the fourth right-hand side term in the last expression, is far more complex. This part can be rewritten as $(\sum_{k \in \mathcal{N}^1} \theta^{k1} - \sum_{k \in \mathcal{X}^0} \theta^{k0})$ times a mean covariance (of size and productivity level). It is unclear how this could be interpreted.

2.9 The Choice of Weights

The question which weights θ^{kt} are appropriate when a choice has been made as to the productivity levels $PROD^{kt}$ ($k \in \mathcal{K}^t$) has received some attention in the literature. Given that somehow $PROD^{kt}$ is output divided by input, should θ^{kt} be output- or input-based? And how is this related to the type of mean—arithmetic, geometric, or

harmonic? The literature does not provide us with definitive answers.¹⁷ Indeed, as long as one stays in the bottom-up framework it is unlikely that a convincing answer can be obtained. We need the complementary top-down view.

A bit formally, the problem can be posed as follows. Generalizing the three definitions used in Sects. 2.5 and 2.6, aggregate productivity is a weighted ‘mean’ of individual productivities

$$PROD^t \equiv M(\theta^{kt}, PROD^{kt}; k \in \mathcal{K}^t), \quad (2.63)$$

where the ‘mean’ $M(\cdot)$ can be arithmetic, geometric, or harmonic; the weights θ^{kt} may or may not add up to 1; and $PROD^{kt}$ can be value added based $TFPROD$, $LPROD$ or $SLPROD$, as defined in Sect. 2.4.1, or gross-output based $TFPROD$ or $SLPROD$, as defined in the extended paper. The task then is: find the set of weights such that

$$PROD^t = PROD^{\mathcal{K}^t}; \quad (2.64)$$

that is, such that aggregate productivity can be interpreted as productivity of the aggregate.

It is clear that there are a number of options here, but the discussion of these can be found in the extended version of this paper.

2.10 Conclusion

The main lessons can be summarized as follows:

1. Generically, productivity is defined as output over input. Yet most, if not all, empirical studies are not about productivity as such, because there is contamination by price effects at the input and/or at the output side of the production units considered. In many sectoral studies the available deflators are more or less deficient; for instance, value added is single-deflated instead of double-deflated. In almost all microdata studies there are simply no firm- or plant-specific deflators available and higher-level substitutes must be used instead. All this may or may not matter at the aggregate (industry or economy)

¹⁷As Karagiannis (2013) showed, the issue is not unimportant. He considered the OP decomposition (2.60) on Greek cotton farm data. Output and input shares were used to weight total factor productivity and labour productivity levels. The covariances turned out to be significantly different. An earlier example was provided by van Beveren (2012), using firm-level data from the Belgian food and beverage industry. de Loecker and Konings (2006) noted that there is no clear consensus on the appropriate weights (shares) that should be used. In their work they used employment based shares $L^{kt} / \sum_k L^{kt}$ to weight value-added based total factor productivity indices $Q_{VA}^k(t, b) / Q_{KL}^k(t, b)$.

- level, but it does matter when it comes to judging the contribution of specific (sets of) production units to aggregate productivity (change).
2. Economists appear to have a preference for working with levels; e.g. with concepts such as real value added. It is good to realize, as pointed out in Sect. 2.4, that a level actually is a long-term index. And this implies that there is always some, essentially arbitrary, normalization involved. For instance, there is a time period for which real value added equals nominal value added; or, there is a period for which total factor productivity equals 1.
 3. Essentially the bottom-up approach consists in aggregating micro-level productivities with help of some set of size-related weights and then decomposing aggregate productivity change into contributions of (specific sets of) continuing, entering, and exiting units. We have seen that there is a large number of such decompositions available. Because of its symmetry and its natural benchmarks for exiting and entering production units we prefer the Diewert-Fox decomposition.
 4. Beware of the covariance, so-called “reallocation”, terms; e.g. in expressions (2.43), (2.49), or (2.60). They are statistical artefacts and there is not necessarily some underlying economic process involved.
 5. In the bottom-up approach not every combination of micro-level productivities, weights, and aggregator function leads to a nice interpretation of aggregate productivity as productivity of the aggregate. The complementary top-down approach should be our guide here. The connection between the two approaches is discussed in the extended version of this paper.

References

- Abbott TA (1991) Producer price dispersion, real output, and the analysis of production. *J Prod Anal* 2:179–195
- Altomonte C, Nicolini M (2012) Economic integration and the dynamics of firms’ competitive behavior. *Struct Chang Econ Dyn* 23:383–402
- Baily MN, Hulten C, Campbell D (1992) Productivity dynamics in manufacturing plants. *Brook Pap Econ Act Microecon* 2:187–249
- Baily MN, Bartelsman EJ, Haltiwanger J (2001) Labor productivity: structural change and cyclical dynamics. *Rev Econ Stat* 83:420–433
- Baldwin JR, Gu W (2006) Plant turnover and productivity growth in Canadian manufacturing. *Ind Corp Chang* 15:417–465
- Baldwin JR, Gu W (2008) Firm turnover and productivity growth in the Canadian retail trade sector. *Economic Analysis (EA)* research paper series, vol 53. Statistics Canada, Ottawa
- Baldwin JR, Gu W (2011) Firm dynamics and productivity growth: a comparison of the retail trade and manufacturing sectors. *Ind Corp Chang* 20:367–395
- Balk BM (2003) The residual: on monitoring and benchmarking firms, industries, and economies with respect to productivity. *J Prod Anal* 20:5–47
- Balk BM (2008) Price and quantity index numbers: models for measuring aggregate change and difference. Cambridge University Press, New York
- Balk BM (2010) An assumption-free framework for measuring productivity change. *Rev Income Wealth* 56(Special Issue 1):S224–S256

- Balk BM (2014) Dissecting aggregate output and labour productivity change. *J Prod Anal* 42:35–43
- Balk BM (2015) Measuring and relating aggregate and subaggregate total factor productivity change without neoclassical assumptions. *Statistica Neerlandica* 69:21–28
- Balk BM, Hoogenboom-Spijker E (2003) The measurement and decomposition of productivity change: exercises on the Netherlands' manufacturing industry. Discussion Paper 03001. Statistics Netherlands, Den Haag
- Balk BM, Reich UP (2008) Additivity of national accounts reconsidered. *J Econ Soc Meas* 33:165–178
- Bartelsman EJ, Dhrymes PJ (1998) Productivity dynamics: US manufacturing plants, 1972–1986. *J Prod Anal* 9:5–34
- Bartelsman EJ, Doms M (2000) Understanding productivity: lessons from longitudinal microdata. *J Econ Lit* XXXVIII:569–594
- Bartelsman EJ, Wolf Z (2014) Forecasting aggregate productivity using information from firm-level data. *Rev Econ Stat* 96:745–755
- Bartelsman EJ, Haltiwanger J, Scarpetta S (2013) Cross-country differences in productivity: the role of allocation and selection. *Am Econ Rev* 103:305–334
- van Beveren I (2012) Total factor productivity estimation: a practical review. *J Econ Surv* 26:98–128
- van Biesebroeck J (2009) Disaggregate productivity comparisons: sectoral convergence in OECD countries. *J Prod Anal* 32:63–79
- Böckerman P, Maliranta M (2007) The micro-level dynamics of regional productivity growth: the source of divergence in Finland. *Reg Sci Urban Econ* 37: 165–182
- Böckerman P, Maliranta M (2012) Globalization, creative destruction, and labour share change: evidence on the determinants and mechanisms from longitudinal plant-level data. *Oxford Econ Pap* 64: 259–280
- Collard-Wexler A, de Loecker J (2013) Reallocation and technology: evidence from the US steel industry. Working Paper 18739. National Bureau of Economic Research, Cambridge
- De Avillez R (2012) Sectoral contributions to labour productivity growth in Canada: does the choice of decomposition formula matter? *Int Prod Monit* 24:97–117
- Dietzenbacher E, Los B (1998) Structural decomposition techniques: sense and sensitivity. *Econ Syst Res* 10:307–323
- Diewert WE, Fox KJ (2010) On measuring the contribution of entering and exiting firms to aggregate productivity growth. In: Diewert WE, Balk BM, Fixler D, Fox KJ, Nakamura AO (eds) Price and productivity measurement. Index number theory, vol 6. Trafford Press, Victoria. www.vancouvervolumes.com. www.indexmeasures.com. Revised version of Discussion Paper 05–02. Department of Economics, University of British Columbia, Vancouver (2005)
- Dumagan JC (2013) A generalized exactly additive decomposition of aggregate labor productivity growth. *Rev Income Wealth* 59:157–168
- Dumagan JC, Balk BM (2015) Dissecting aggregate output and labour productivity change: a postscript on the role of relative prices. *J Prod Anal* (online)
- Durand R (2004) Uniqueness of the numeraire and consistent valuation in accounting for real values. *J Econ Soc Meas* 29: 411–426
- Eslava M, Haltiwanger J, Kugler A, Kugler M (2013) Trade and market selection: evidence from manufacturing plants in Colombia. *Rev Econ Dyn* 16:135–158
- Foster L, Haltiwanger J, Krizan CJ (2001) Aggregate productivity growth: lessons from microeconomic evidence. In: Hulten CR, Dean ER, Harper MJ (eds) New developments in productivity analysis. Studies in income and wealth, vol 63. The University of Chicago Press, Chicago/London
- Foster L, Haltiwanger J, Krizan CJ (2006) Market selection, reallocation, and restructuring in the US retail trade sector in the 1990s. *Rev Econ Stat* 88:748–758
- Foster L, Haltiwanger J, Syverson C (2008) Reallocation, firm turnover, and efficiency: selection on productivity or profitability? *Am Econ Rev* 98:394–425

- Fox KJ (2012) Problems with (dis)aggregating productivity, and another productivity paradox. *J Prod Anal* 37:249–259
- Griliches Z, Regev H (1995) Firm productivity in Israeli industry, 1979–1988. *J Econ* 65:175–203
- Haltiwanger J (1997) Measuring and analyzing aggregate fluctuations: the importance of building from microeconomic evidence. *Fed Reserve Bank St. Louis Econ Rev* 79(3):55–77
- Hansell D, Nguyen T (2012) Productivity, entry and exit in Australian manufacturing and professional services. Paper presented at economic measurement group workshop, University of New South Wales, Sydney, 21–23 November 2012
- Hsieh CT, Klenow PJ (2009) Misallocation and manufacturing TFP in China and India. *Q J Econ* 124:1403–1448
- Hytyinen A, Maliranta M (2013) Firm lifecycles and evolution of industry productivity. *Res Policy* 42:1080–1098
- Karagiannis G (2013) Reallocation and productivity dynamics: empirical evidence on the role of aggregation weights. Mimeo. Department of Economics, University of Macedonia, Thessaloniki
- Kauhanen A, Maliranta M (2012) Micro-components of aggregate wage dynamics. ETLA Working papers No. 1. The Research Institute of the Finnish Economy.
- Kirwan BE, Uchida S, White TK (2012) Aggregate and farm-level productivity growth in tobacco: before and after the quota buyout. *Am J Agric Econ* 94:838–853
- Lin YC, Huang TH (2012) Creative destruction over the business cycle: a stochastic frontier analysis. *J Prod Anal* 38:285–302
- de Loecker J, Konings J (2006) Job reallocation and productivity growth in a post-socialist economy: evidence from Slovenian manufacturing. *Eur J Polit Econ* 22:388–408
- Mairesse J, Jaumandreu J (2005) Panel-data estimates of the production function and the revenue function: what difference does it make? *Scand J Econ* 107:651–672
- Marin AG, Voigtlander N (2013) Exporting and plant-level efficiency gains: it's in the measure. Working Paper 19033. National Bureau of Economic Research, Cambridge
- Melitz MJ, Polanec S (2015) Dynamic Olley-Pakes productivity decomposition with entry and exit. *Rand J Econ* 46: 362–375
- Nishida M, Petrin A, Polanec S (2014) Exploring reallocation's apparent weak contribution to growth. *J Prod Anal* 42:187–210
- OECD (2014) The micro drivers of aggregate productivity. Report DSTI/EAS/IND/WPIA(2014)4. Organisation for Economic Co-operation and Development, Paris
- Olley S, Pakes A (1996) The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64:1263–1297
- Reinsdorf M (2015) Measuring industry contributions to labour productivity change: a new formula in a chained Fisher index framework. *Int Prod Monit* 28: 3–26
- Smeets V, Warzynski F (2013) Estimating productivity with multi-product firms, pricing heterogeneity and the role of international trade. *J Int Econ* 90:237–244
- Syverson C (2011) What determines productivity? *J Econ Lit* 49:326–365
- Tang J, Wang W (2004) Sources of aggregate labour productivity growth in Canada and the United States. *Can J Econ* 37:421–444
- Vainiomäki J (1999) Technology and skill upgrading: results from linked worker-plant data for Finnish manufacturing. In: Haltiwanger JC, Lane JI, Spletzer JR, Theeuwes JHM, Troske KR (ed) The creation and analysis of employer-employee matched data. Contributions to Economic Analysis 241. Emerald Bingley
- Wolf Z (2011) Aggregate productivity growth under the microscope. Research Series, vol 518. Tinbergen Institute, VU University Amsterdam

Chapter 3

A General Error Revenue Function Model with Technical Inefficiency: An Application to Norwegian Fishing Trawler

Subal C. Kumbhakar, Frank Asche, Kristin Roll, and Ragnar Tveteras

Abstract In this paper we consider a revenue maximizing model and derive the revenue function from the transformation function where errors are parts of the outputs. First, we adapt McElroy's additive general error model to the transformation function with multiple inputs and multiple outputs and derive the revenue function. The error terms in the output supply functions, derived from the revenue function, inherit their stochasticity from the error terms in the outputs. The second approach uses a multiplicative general error model (MGEM), in the spirit of Kumbhakar and Tsionas (J Appl Econ 26:270–297, 2011), with multiple outputs in which multiplicative errors are parts of the outputs. The MGEM is further generalized to accommodate output-oriented inefficiency. The translog revenue function with MGEM makes the intercept and the coefficients of the linear terms random (functions of the errors associated with outputs). The errors in the revenue share functions are linear functions of the errors associated with the outputs. Vessel level data for the Norwegian whitefish fisheries for the period 1995–2007 is used to showcase the application of the model. We also estimate a standard (off-the-shelf) revenue function with output-oriented technical inefficiency and compare technical change and technical efficiency results with the MGEM revenue function which is estimated along with the revenue share equations. Although the means are found to be somewhat similar, patterns of technical change and technical efficiency are found to be quite different across these models.

S.C. Kumbhakar (✉)

Department of Economics, State University of New York, Binghamton, NY 13902, USA

University of Stavanger Business School, N-4036 Stavanger, Norway

e-mail: kkar@binghamton.edu

F. Asche

Department of Industrial Economics, University of Stavanger, N-4036 Stavanger, Norway

K. Roll

University of Stavanger, N-4036 Stavanger, Norway

R. Tveteras

University of Stavanger Business School, N-4036 Stavanger, Norway

Keywords Transformation function • Multiple outputs • Technical inefficiency • Fisheries

JEL Classification No.: C51, D24, Q22

3.1 Introduction

In the econometric estimation of production, cost, revenue and profit functions the specification of error terms is crucial. The standard neoclassical approach assumes these functions to be deterministic. Some researchers (Brown and Walker 1995; McElroy 1987; Chavas and Segerson 1987; Anderson and Blundell 1982; among others) challenged this view and demonstrated that estimates and inferences drawn from these models depend crucially on the underlying stochastic process. The stochastic nature of the empirical model should come from the theory and the random variables in the empirical model should be subject to the same theoretical restrictions that apply to the deterministic parts of the models. In most of the empirical work well-behaved error terms are appended to the estimating equations that are derived from a deterministic model. This simplistic approach may very well be in conflict with a coherent theoretical model. For example, in a cost minimizing model, McElroy (1987) showed that input demand equations with additive homoskedastic errors result in cost share equations with complicated error terms. She showed that this problem can be avoided by using an additive error structure in estimating the conditional demand functions.

In this paper we start from a multi-output transformation function in which the error terms appear either additively or multiplicatively with the outputs. We consider a revenue maximizing model in which the revenue function inherits the error terms from the outputs in the transformation function. First, we adapt McElroy's additive general error model (AGEM) to the transformation function and derive the revenue function. The error terms in the output supply functions, derived from the revenue function, inherit their stochasticity from the error terms from the transformation function. The second approach uses a multiplicative general error model (MGEM) in the spirit of Kumbhakar and Tsionas (2011). The MGEM is generalized to accommodate multiple outputs. The resulting revenue and revenue share functions inherit all the errors from the outputs in the transformation function. The paper departs from the traditional modeling of revenue function in which both noise and inefficiency terms are added in an ad hoc fashion. The model is neatly structured in the sense that nothing is added in an ad hoc fashion. First we consider the model without inefficiency. For modeling inefficiency we assume that inefficiency is multiplicatively related to the errors associated with each output.

Similar to McElroy, who used a cost minimizing behavior, we find that the output supply system is easier to estimate for the AGEM formulation. On the other hand, it is easier to estimate the revenue share system under the MGEM, which is what we focus on empirically. The error terms under both specifications are shown to

be related to technical inefficiency. More specifically, we show that the error in the revenue function has a complicated form and is related to pure noise as well as technical inefficiency.

Little attention has been given in the literature to the possibility of inefficient output mix in multi-output firms. This is of particular interest in industries where effort is exogenous in the short run, while output levels are endogenous. One such industry is fisheries. It is well known that firm performance vary, which is often attributed to managerial skills or so-called skipper effects (Kirkley et al. 1998). There is a substantial literature investigating technical efficiency in fisheries, but these studies use traditional primal approaches. Some recent examples are Gutormsen and Roll (2011), Pascoe et al. (2012) and Squires and Vestergaard (2013). When investigating firm behavior but not investigating inefficiency, revenue functions are commonly used as the effort in form of a vessel with a given crew size is taken as exogenous (Kirkley and Strand 1988; Squires and Kirkley 1991; Campbell and Nicholl 1994; Thunberg et al. 1995; Diop and Kazmierczak 1996; Asche 2009). Our approach is ideal to investigate whether the output mix is sub-optimal, a feature that seems likely given the presence of skipper effects.

We use vessel level data for the Norwegian whitefish fisheries for the period 1995–2007. For each vessel we observe the key vessel characteristics such as length, tonnage, crew size and days at sea. The vessels span a range from small coastal vessels at less than 10 m to large trawlers, and the fleet uses a variety of gears like nets, long-lines, Danish seines and trawls. Thus, the inputs are the capital (vessel), labor, and days at sea. The outputs are different species caught: cod, saithe and haddock, and the other species (mainly bycatch that we treat as a single specie). Gutormsen and Roll (2011) found substantial inefficiency in this fleet using a production function. They did not account for the multi-output nature of the fishery but with results suggesting that this is an issue of potential importance.

The rest of the paper is structured as follows. Section 3.2 outlines the general error models (both additive and multiplicative) and derives the revenue function which contains errors associated with the outputs. Inefficiency in the multiplicative error model is introduced in Sect. 3.3. Section 3.4 describes the data and discusses the related literature. Estimation of the models is described in Sect. 3.5 and result are discussed in Sect. 3.6. Section 3.7 concludes the paper.

3.2 The General Error Models

3.2.1 The General Additive Error Model

McElroy (1987) proposed an additive general error model (AGEM) for the production function and derived the stochastic specifications for the derived demand and cost systems. The production function with AGEM is written as

$$y = f(x_1 - \varepsilon_1, x_2 - \varepsilon_2, \dots, x_J - \varepsilon_J), \quad x_j - \varepsilon_j > 0 \quad \forall j \quad (3.1)$$

The error vector ε is assumed to be known to the firm but unknown to the researcher. Furthermore the elements of ε are small enough so that $x_j - \varepsilon_j > 0 \forall j$. These are given a classical measurement error interpretation. The advantage of this specification is that no additional errors are needed for the dual cost function and cost share equations, as well as the input demand system.

The above model can be easily generalized to accommodate multiple outputs in which the production technology is specified in terms of a transformation function $F(y - \theta, x, t) = 0$, where y is vector of M outputs. In this specification inputs are exogenous and this is why no errors are added to the input vector. On the other hand, errors are allowed in the outputs which are endogenous. The elements of θ are assumed to be known to the firm but unknown to the researcher. Furthermore the elements of θ are small enough so that $y_m - \theta_m > 0 \forall m$. Similar to McElroy these θ_m are given a classical measurement error interpretation.

Assume that fisheries maximize revenue. The revenue maximization problem is

$$\text{Max} \sum_m p_m y_m \text{ subject to } F(y - \theta, x, t) = 0 \quad (3.2)$$

which gives the following first-order conditions (FOCs)

$$p_m + \lambda F_m(y - \theta, x, t) = 0, m = 1, \dots, M \quad (3.3)$$

where λ is the Lagrange multiplier and p_m is the price of output y_m . Denote $y_m^* \equiv y_m - \theta_m$ and use the above FOCs along with the transformation function $F(y^*, x, t) = 0$ to solve for y_m^* , viz., $y_m^* = \psi_m(p, x, t) \Rightarrow y_m = \theta_m + \psi_m(p, x, t)$. Thus, θ_m is transmitted linearly to the output supply function. We call the solution of y_m^* the output supply function because $y_m = y_m^* + \theta_m$ where $y_m^* = \psi_m(p, x, t)$ is derived from the Envelope theorem (Hotelling's lemma, i.e., $\frac{\partial R^*}{\partial p_m} = y_m^*$, once a parametric functional for R^* is chosen). Using these output supply functions, optimum (maximum) revenue is obtained from $\sum_m p_m y_m^* = \sum_m p_m \psi_m(p, x, t) = R^*(p, x, t)$. Note that R^* is unobserved. Observed revenue $R = \sum_m p_m y_m$ can be related to R^* from $R = \sum_m p_m y_m = R^*(p, x, t) + \sum_m p_m \theta_m$.

Given that the error structure is additive it is simpler to estimate the system of output supply equations $y_m = \psi_m(p, x, t) + \theta_m$ starting from a parametric form of $R^*(p, x, t)$. There is no need to estimate the revenue function per se. Further, there is no need to add any stochastic terms to the supply functions.

3.2.2 The Multiplicative General Error Model

There are some issues with the specification of the functional form for $R^*(.)$. Note that R^* is homogenous of degree 1 in p . The linear homogeneity restrictions cannot be imposed, for example, on a quadratic function. Specifying a normalized quadratic revenue function (normalized by one of the output prices) might help but it creates

other problems, viz., invariance of it to the choice of the normalizing output price. There is another problem. The multi-output AGEM is not suitable for modeling technical inefficiency.

Because of these problems we consider an alternative modeling approach. It is often preferred to estimate the model in logarithm to accommodate heterogeneous units (firms) in the data. Although normalizing the data at the mean (median) helps, the model in which the covariates appear in log can accommodate the heterogeneity issue better. The log model can also impose the linear homogeneity (in output prices) restrictions easily. Further, technical inefficiency in the log model has a nice interpretation (percentage under-production of outputs or over-use of inputs depending on whether an output or input oriented inefficiency is used). Thus when it comes to a choice between the quadratic and the translog, the latter is always preferred.

In view of this, we propose using a multiplicative general error formulation and derive the corresponding revenue function and the revenue share functions. Since inputs in a revenue maximization model are treated as exogenous we specify the transformation function as

$$F(\theta \odot y, x, t) \equiv F(y^*, x, t) = 1 \quad (3.4)$$

where \odot represents Hadamard product (element-wise multiplication) and $\theta_m > 0$ is the error term associated with output y_m . The revenue maximization problem is: Max $\sum_m p_m y_m$ subject to $F(y^*, x, t) = 1$.

The FOCs of the above problem are:

$$\frac{p_m}{p_1} = \frac{F_m(\cdot)}{F_1(\cdot)} \frac{\theta_m}{\theta_1}, m = 2, \dots, M \implies \frac{p_m}{p_1} \div \frac{\theta_m}{\theta_1} \equiv \frac{p_m^*}{p_1^*} = \frac{F_m(\cdot)}{F_1} \quad (3.5)$$

where $p_m^* = p_m / \theta_m, m = 1, \dots, M$.

The above $(M - 1)$ FOCs in (3.5) along with the transformation function in (3.4) can be solved for $y_m^* = \psi_m(p, x, t), m = 1, \dots, M$. We use these solutions to define the pseudo revenue function $R^*(p, x, t) = \sum_m p_m^* \psi_m(p, x, t)$. The advantage of defining R^* is that we can use the envelope theorem (equivalent of Hotelling's lemma) to derive the pseudo output supply functions, $y_m^* = \frac{\partial R^*}{\partial p_m^*}$. Furthermore, we can relate the unobserved R^* to the observed revenue $R = \sum_m p_m y_m = \sum_m p_m^* y_m^* = R^*$. That is, $R^* = R$.

Here the starting point is a logarithmic form of R^* , i.e., a translog form of $\ln R^*$, where we use the Envelope theorem to derive the revenue share functions. These are $\partial \ln R^* / \partial \ln p_m^* = p_m^* y_m^* / R^* = p_m y_m / R \equiv RS_m, m = 2, \dots, M$. The $(M - 1)$ revenue share (RS_m) equations and the revenue function makes it a complete system with M equations and M endogenous outputs. The error terms in these equations are built into the system through the elements of θ . There is no need to add any extra error term in any of the equations.

If the translog revenue function is expressed as

$$\begin{aligned} \ln R = & \beta_0 + \sum_m \beta_m \ln p_m^* + \sum_j \alpha_j \ln x_j + \alpha_t t \\ & + .5 \sum_m \sum_n \beta_{mn} \ln p_m^* \ln p_n^* + \sum_m \sum_j \gamma_{mj} \ln p_m^* \ln x_j + \sum_m \alpha_{mt} \ln p_m^* t \\ & + .5 \sum_j \sum_k \delta_{jk} \ln x_j \ln x_k + \sum_j \eta_{jt} \ln x_j t + .5 \alpha_{tt} t^2 \end{aligned} \quad (3.6)$$

the corresponding revenue share equations are

$$RS_m = \beta_m + \sum_n \beta_{mn} \ln p_n + \sum_j \gamma_{mj} \ln x_j + \alpha_{mt} t + \zeta_m \quad (3.7)$$

where $\zeta_m = -\sum_n \beta_{mn} \ln \theta_n$ can be viewed as the error term in the m th revenue share equation. We used $\ln p_m^* = \ln p_m - \ln \theta_m$ to get ζ_m in the above revenue share equation.

We can do similar substitution in the revenue function to get

$$\begin{aligned} \ln R = & [\beta_0 - \sum_m \beta_m \ln \theta_m + .5 \sum_m \sum_n \beta_{mn} \ln \theta_m \ln \theta_n] \\ & + \sum_m [\beta_m - \sum_n \beta_{mn} \ln \theta_m] \ln p_m \\ & + \sum_j [\alpha_j - \sum_j \gamma_{mj} \ln \theta_m] \ln x_j + [\alpha_t - \sum_m \alpha_{mt} \ln \theta_m] t \\ & + .5 \sum_m \sum_n \beta_{mn} \ln p_m \ln p_n + \sum_m \sum_j \gamma_{mj} \ln p_m \ln x_j + \sum_m \alpha_{mt} \ln p_m t \\ & + .5 \sum_j \sum_k \delta_{jk} \ln x_j \ln x_k + \sum_j \eta_{jt} \ln x_j t + .5 \alpha_{tt} t^2 \end{aligned} \quad (3.8)$$

which makes the intercept and the coefficients of $\ln x_j$, $\ln p_m$ and t random.

If one starts from a standard neoclassical transformation function $F(y, x, t) = 1$, the resulting translog revenue function will be of the form

$$\begin{aligned} \ln R = & \beta_0 + \sum_m \beta_m \ln p_m + \sum_j \alpha_j \ln x_j + \alpha_t t \\ & + .5 \sum_m \sum_n \beta_{mn} \ln p_m \ln p_n + \sum_m \sum_j \gamma_{mj} \ln p_m \ln x_j + \sum_m \alpha_{mt} \ln p_m t \\ & + .5 \sum_j \sum_k \delta_{jk} \ln x_j \ln x_k + \sum_j \eta_{jt} \ln x_j t + .5 \alpha_{tt} t^2 + \psi \end{aligned} \quad (3.9)$$

when a zero mean and constant variance random term (ψ) is added in an ad hoc fashion. Note that all the coefficients in (3.9) are constants. If we rewrite (3.8) in the form of (3.9) the error term will be $v = -[\sum_m \beta_m \ln \theta_m + \sum_m \sum_n \beta_{mn} \ln \theta_m \ln p_n + \sum_m \sum_j \gamma_{mj} \ln \theta_m \ln x_j + \sum_m \alpha_{mt} \ln \theta_m t] + .5 \sum_m \sum_n \beta_{mn} \ln \theta_m \ln \theta_n$, which is heteroscedastic and has a non-zero mean. So the main difference between the two modeling approaches is in terms random versus non-random coefficients.

3.3 Modeling Inefficiency

3.3.1 Inefficiency in MGEM

Because of the difficulty in imposing linear homogeneity restrictions on the revenue function in the AGEM, from now on we focus on the MGEM. We start with the same transformation function $F(\theta \odot y, x, t) \equiv F(y^*, x, t) = 1$ where $\theta_m = \theta \cdot v_m$; $\theta \geq 1$ captures output-oriented (radial) technical efficiency and $v_m > 0$ is output-specific noise (measurement error).

If the revenue function is translog, it can be expressed as

$$\begin{aligned} \ln R = & \beta_0 + \sum_m \beta_m \ln p_m^* + \sum_j \alpha_j \ln x_j + \alpha_t t \\ & + .5 \sum_m \sum_n \beta_{mn} \ln p_m^* \ln p_n^* + \sum_m \sum_j \gamma_{mj} \ln p_m^* \ln x_j + \sum_m \alpha_{mt} \ln p_m^* t \\ & + .5 \sum_j \sum_k \delta_{jk} \ln x_j \ln x_k + \sum_j \eta_{jt} \ln x_j t + .5 \alpha_{tt} t^2 \end{aligned} \quad (3.10)$$

The corresponding revenue share equations are

$$RS_m = \beta_m + \sum_n \beta_{mn} \ln p_n^* + \sum_j \gamma_{mj} \ln x_j + \alpha_{mt} t \quad (3.11)$$

Note that no ad hoc error terms are added to any of the equations above. Since $\ln p_m^* = \ln p_m - \ln \theta_m = \ln p_m - \ln v_m - \ln \theta$, we can express the revenue share equations above as

$$RS_m = \beta_m + \sum_n \beta_{mn} \ln p_n + \sum_j \gamma_{mj} \ln x_j + \alpha_{mt} t + \zeta_m + \ln \theta \sum_n \beta_{mn} \quad (3.12)$$

where $\zeta_m = -\sum_n \beta_{mn} \ln v_n$.

We rewrite the revenue function and the revenue share equations after imposing the linear homogeneity (in prices) constraints, viz., $\sum_m \beta_m = 1$, $\sum_n \beta_{mn} = 0 \forall m$, $\sum_m \gamma_{mj} = 0 \forall j$, $\sum_m \alpha_{mt} = 0$. This is done by normalizing the revenue and

output prices by one of the prices (p_1). With this the above revenue function and the associated share equations are:

$$\begin{aligned}
 \ln(R/p_1) = & \beta_0 + \sum_m \beta_m (\ln \tilde{p}_m - \ln \mu_m) + \sum_j \alpha_j \ln x_j + \alpha_t t \\
 & + .5 \sum_m \sum_n \beta_{mn} (\ln \tilde{p}_m - \ln \mu_m) (\ln \tilde{p}_n - \ln \mu_n) \\
 & + \sum_m \sum_j \gamma_{mj} (\ln \tilde{p}_m - \ln \mu_m) \ln x_j \\
 & + \sum_m \alpha_{mt} (\ln \tilde{p}_m - \ln \mu_m) t \\
 & + .5 \sum_j \sum_k \delta_{jk} \ln x_j \ln x_k + \sum_j \eta_{jt} \ln x_j t + .5 \alpha_{tt} t^2 + v_1 - \ln \theta
 \end{aligned} \tag{3.13}$$

$$RS_m = \beta_m + \sum_n \beta_{mn} \ln \tilde{p}_n + \sum_j \gamma_{mj} \ln x_j + \alpha_{mt} t + \epsilon_m, m = 2, \dots, M, \tag{3.14}$$

where $\tilde{p}_m = (p_m/p_1)$, $\mu_m = (\theta_m/\theta_1) \equiv (v_m/v_1)$ and $\epsilon_m = -\sum_{n=2} \beta_{mn} \ln(v_n/v_1)$, $m = 2, \dots, M$.

We can rewrite the $(M - 1)$ revenue share equations in (3.14) in matrix form,

$$RS = \beta + B \ln \tilde{p} + \Gamma \ln x + a_t t + \epsilon \tag{3.15}$$

where β and a_t are column vectors of order $(M - 1)$ whose elements are β_m and α_{mt} respectively, and B , Γ are matrices of order $(M - 1) \times (M - 1)$ and $(M - 1) \times K$ whose elements are $\{\beta_{mn}\}$ and $\{\gamma_{mj}\}$.

3.3.2 Inefficiency in the Standard Revenue Function

For this we write the transformation function with output-oriented efficiency as $F(\tilde{y}, x, t) = 1$ where $\tilde{y} = \theta \cdot y$, and $\theta \geq 1$ (a scalar) represents output-oriented technical efficiency. The revenue maximization problem is

$$\text{Max} \sum_m p_m y_m \text{ subject to } F(\tilde{y}, x, t) = 1 \tag{3.16}$$

gives the following FOCs:

$$p_m/p_1 = F_m(\cdot)/F_1(\cdot) = 0, m = 2, \dots, M \tag{3.17}$$

which along with the transformation function $F(\tilde{y}, x, t)$ can be used to solve for, viz., $\tilde{y}_m = \psi_m(p, x, t)$. Use the solution of \tilde{y}_m —the output supply function—to define a pseudo revenue function $\tilde{R} = \sum_m p_m \tilde{y}_m(p, x, t)$. The Envelope theorem (Hotelling's lemma) can then be used to derive $\tilde{y}_m(p, x, t)$ from $\frac{\partial \tilde{R}}{\partial p_m}$. The observed revenue $R = \sum_m p_m y_m$ can be related to \tilde{R} from $R(.) = \sum_m p_m \tilde{y}_m(p, x, t) = \theta[\sum_m p_m y_m(.)] = \theta.R \Rightarrow \ln R = \ln \tilde{R}(.) - \ln \theta$. Note that there is no error term in the above revenue function. A classical error term is appended before estimating the revenue function.

One can use the relationship $\ln R = \ln \tilde{R} - \ln \theta$ and specify $\ln \tilde{R}$ by a translog function as in (3.9) to derive the revenue share equations $RS_m = \frac{\partial \ln \tilde{R}}{\partial \ln p_m}$. Note that these share equations are without any errors. If error terms are added to these share equations in an ad hoc fashion and are interpreted as optimization errors, these errors are likely to affect output supply and revenue. In other words, the revenue function will also be a function of these optimization errors. For example, since the revenue shares are $RS_m = \frac{\partial \ln \tilde{R}}{\partial \ln p_m} + \eta_m$ and the revenue function has to satisfy the integrability condition, it should be of the form $\ln R = \ln \tilde{R} + \sum_m \ln p_m \eta_m + v - u$ where v is the classical error term added in an ad hoc fashion to the revenue function, and $u = \ln \theta$. So the only way one can justify the revenue function specification $\ln R = \ln \tilde{R} + v - u$ is to assume that there are no optimization errors which makes the revenue shares non-stochastic.

3.4 Fisheries Studies and Data

Revenue maximization is a common way to characterize producers' behavior in the fishery economics literature (Carlson 1973; Kirkley and Strand 1988; Squires and Kirkley 1991; Campbell and Nicholl 1994; Thunberg et al. 1995; Diop and Kazmierczak 1996; Asche 2009). Carlson (1973), for instance, suggests that a fishing firm is likely to maximize revenues once the target stocks or geographic region has been specified. Fishermen select species given vessel, regulatory and environmental constraints. Furthermore, fisheries are generally multiproduct, and the inputs are largely fixed and proportional to the size of the vessel. By the time a captain has decided when to make a fishing trip, the crew size, gear type, and likely fishing areas have already been decided. Thus, the captain's options for changing input levels can be severely limited. Consequently, revenue maximization subject to given inputs (which are predetermined) appears to be a reasonable assumption for a multispecies fishing firm. This is therefore the preferred approach when high frequency data such as trip data is used, even when the fishery is regulated with an annual vessel quota.

To illustrate the approach we use a panel data from the Norwegian whitefish fisheries. Several studies have investigated productivity for this fleet under different regulatory scenarios (Bjørndal and Gordon 1993; Salvanes and Squires 1995; Asche et al. 2009; Asche 2009; Guttormsen and Roll 2011; Kumbhakar et al. 2013) using annual data.

In this paper we use a panel of trip data for Norwegian whitefish trawlers. The Norwegian whitefish fisheries are by far the most valuable fisheries in Norway. It is year round fisheries with substantial seasonality in landings (Asche 2009). Cod, haddock and saithe are the main species targeted, while a number of other species are caught primarily as bycatch. A large part of the fishery is from stocks that are shared with other countries, and Norway has bilateral agreements with, Russia, EU, Iceland, Greenland and the Faroe Islands. Due to a dramatic decline in the stock in Norwegian Arctic Cod, quota regulations commenced in 1982, first only with a Total Allowable Catch (TAC) for the fleet, and from 1996 with individual vessel quotas. The fishery is regulated with individual vessel quotas for cod, haddock and saithe, with TAC for some of the remaining species and no regulations for the majority of the remaining species.

The data in this study is based on trip data for an unbalanced panel of Norwegian vessels observed for the years 1995–2007. Total number of observations is 747. The data are provided by the Norwegian Directorate of Fisheries, which has collected the data for several sources. Landing records provide quantity and value landed as well as trip length for each trip. Vessel characteristics are collected in an annual survey.

The data contains information on type of gear, age of the vessel, yearly fuel consumption, taxes, insurance costs, maintenance, depreciation, total wage, size of crew, tonnage, horsepower, length of the vessel and type of vessel. The input variables are: capital (measured by tonnage of the vessel), x_k ; labor (measured by number of man days), x_l ; fuel measured by (real) fuel expenditure, x_f . The output variables are: cod (y_2), haddock (y_3), saithe (y_4) and a group containing all other species (y_1), all in kilograms. Price per kilogram of each output is defined as total revenue from each species divided by the corresponding output quantity. The time trend variable is introduced to capture technical change. We also used number of operation days as a determinant of inefficiency.

3.5 Estimation

3.5.1 A Single Equation Approach

The revenue function that is widely used in practice to estimate inefficiency is specified as $\ln R = \ln \tilde{R} + v - u$ where ψ is the noise term added *ex post* in an ad hoc and u is output-oriented (radial) inefficiency. To estimate this model (labeled as Model 0) we assume (1) a translog function for $\ln \tilde{R}$ (as in (3.9)), (2) $\psi \sim i.i.d.N(0, \sigma_\psi^2)$, (3) $u \sim N^+(0, \sigma_u^2(z))$, and (4) u and ψ are independent of each other and also independent of the covariates of the revenue function. The variables in z are determinants of inefficiency and $\sigma_u^2(z)$ is usually specified as $\sigma_u^2(z) = \exp(\gamma' z)$ which ensures positivity of $\sigma_u^2(z)$. Note that in this specification the z variables explain inefficiency though the variance of u . If an increase in z increases $\sigma_u^2(z)$

Table 3.1 Summary statistics

Variable	Mean	Std. Dev.
Length	47.5169	10.9976
Tonnage	966.2838	698.2832
Age	17.7309	9.5086
Fuel expenditure	3,386,428	2,995,628
Fuel	280580.5	230,207
Labor cost	8,023,372	4,704,517
Labor (man days)	5930.79	2691.80
Operation days	279.7564	61.9517
Cod price	11.57378	3.295626
Cod quantity	940283.6	606233.4
Haddock price	8.810501	3.234454
Haddock quantity	340184.3	262008.3
Saithe price	4.940141	1.257026
Saithe quantity	1,187,699	898752.5
Saithe price	4.940141	1.257026
Other fish quantity	594291.6	660931.6
Other fish price	9.433167	3.61343
Revenue	2,530,464	1,608,946
Revenue share cod	0.4312	0.2096
Revenue share haddock	0.1090	0.0467
Revenue share saithe	0.2523	0.1796
Revenue share other	0.2075	0.1691
No. of observations	747	
Year	1995–2007	

the mean inefficiency will increase. Given the above assumptions the model is a similar to a standard production frontier model in which the composed error term is $v - u$. The model can be estimated using the ML method that is standard in the SF literature and inefficiency can be estimated using the Jondrow et al. (1982) formula. The marginal effects of z is obtained using the formula in Wang (2002) which computes the marginal effects by taking partial derivative of $E(u)$ with respect to the z variables (Table 3.1).

3.5.2 Estimation of the MGEM Revenue Model

Since the MGEM is built on the notion of one error for each output and outputs are endogenous, it is desirable to use a system approach to estimate the MGEM. A system approach with inefficiency is somewhat complicated to estimate and there are no publicly/commercially available computer packages to estimate a system. Note that the error term in the revenue function is a complicated function of all the

errors in outputs plus the inefficiency term which appears linearly. As a result of this, derivation of the log likelihood function (starting from distributional assumptions on v_1, \dots, v_m and u) for the system consisting of the revenue function and the revenue share equations is nontrivial. However, since inefficiency is not transmitted to the revenue share equations, it is possible to use a two-step approach which is easier to estimate. These steps are as follows.

Step 1: Estimate the revenue share system using seemingly unrelated regression (SUR) procedure.

$$RS = \beta + B \ln \tilde{p} + \Gamma \ln x + a_t t + \epsilon \quad (3.18)$$

SUR will give consistent estimate of the parameters in β, B, Γ and a_t without any distributional assumptions on $\epsilon_m = -\sum_n \beta_{mn} \ln(v_n/v_1), m = 2, \dots, M$.

Residuals from the share equations can be viewed as an estimate of $\epsilon = -B \ln \tilde{v}$ where the elements of \tilde{v} are (v_n/v_1) . From these residuals we get $\ln \tilde{v} = -B^{-1}\epsilon$.

Step 2: Use the estimated $\ln \tilde{v}$ to define $\ln \tilde{p}^* = \ln \tilde{p} - \ln \mu \equiv \ln \tilde{p} - \ln \tilde{v}$ which can be treated as data in the revenue function (3.13) to estimate its parameters and technical inefficiency $\ln \theta \geq 0$ using the standard stochastic frontier (SF) approach. For this we need to make distributional assumptions on v_1 and $u = \ln \theta$. As before we assume (1) $v_1 \sim i.i.d.N(0, \sigma_1^2)$, (2) $u \sim N^+(0, \sigma_u^2(z))$, and (3) u and v are independent of each other and also independent of the covariates of the revenue function. Based on these assumptions, the revenue function in (3.13) (labeled as Model 1) can be estimated using the standard SF softwares. As in Model 0, inefficiency in Model 1 can also be estimated using the Jondrow et al. (1982) estimator. Note the difference between Model 0 and Model 1. The random coefficients in (3.8) are adjusted by redefining the variables that appear linearly in (3.13) using the estimate of $\ln \mu_m = \ln(\tilde{v}_m)$ from the revenue share equations.

3.6 Results

Since coefficients of the linear terms (after imposing linear homogeneity restrictions) in Model 1 are random whereas the coefficients in Model 0 are constants, one can perform a likelihood ratio test to determine whether Model 1 is supported by the data or not. The test is whether $\mu_m = 1, m = 2, \dots, M$. Without inefficiency this gives Model 0 from (13).¹ The test (random versus fixed coefficients) favors the MGEM formulation and rejects the neoclassical revenue function formulation at the 5 % level of significance.

¹Note that this however does not address the philosophical question about the source of the error terms. If the null is accepted, the share equations will be deterministic.

We focus on two metrics from Model 0 and Model 1. These are technical change (TC) and technical efficiency (TE). The variables used to explain efficiency are logarithm of number of operation days (z) and time trend (t). The temporal behavior of efficiency is estimated from TE change which is nothing but the marginal effect of t , i.e., $TEC = \frac{\partial \ln TE}{\partial t} = -\frac{\partial u}{\partial t}$. We also estimate TC from $TC = \frac{\partial \ln R}{\partial t}$. Thus, TC measures the rate at which revenue changes over time, ceteris paribus. Consequently, a positive (negative) value of TC can be viewed as technical progress (regress)—a shift in the production possibility frontier over time holding everything else unchanged. TE is estimated from the conditional mean of $\exp(-u)$, i.e., $E(\exp(-u)|(v_1 - u))$ in Model 1 and $E(\exp(-u)|(\psi - u))$ in Model 0. Since u and therefore TE depends on z and t , we also compute their marginal effects on TE. The marginal effect of t is the catch-up factor and shows how efficiency changed over time, all else unchanged. On the other hand, the marginal effect of z is the percentage change in efficiency for a 1 % change in number of operation days.

We report the kernel density of estimated TC from Models 0 and 1 (denoted by TC_0 and TC_1 in Fig. 3.1). The distribution of TC_0 is bi-modal with large variations. Both the models show technical regress (negative TC)² for majority of fisheries in most of the years, although their mean values are quite different.

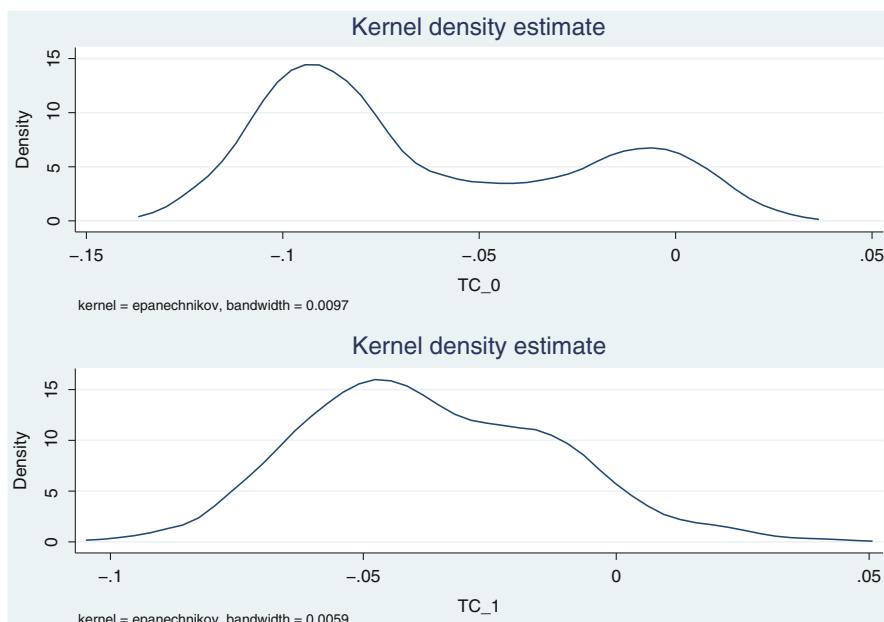


Fig. 3.1 Technical change in Models 0 and 1

²Given that technology improves over time, it would be more plausible to suggest that the negative value is actually zero minus sampling variability. We owe this insight to an anonymous referee.

The mean TC from Models 0 and 1 are -6.34 and -3.67% , respectively. Since the kernel density is based on the pooled data (all firms and years combined), a comparison across models in terms of the kernel densities might not be interesting. To get a measure of closeness of estimates of TC from the two models, we computed the concordance correlation. It is found to be 0.521 with a standard error of 0.017. Thus we find significant positive correlation of TC from the two models but the correlation is far from unity.

We also estimated the standard neoclassical revenue function without inefficiency. The estimated TC from this model is very close to Model 0—the mean of TC being -0.069 and -0.063 . Also the concordance correlation is very high (0.979 with a standard error of 0.001). Since the neoclassical model is rejected against both Model 0 and Model 1 (without inefficiency), we are not reporting results from the neoclassical revenue function.

To examine the differences in the estimated TC in Models 0 and 1 over time, we do a scatter plot of TC from both models and report them in Fig. 3.2. It can be seen from scatter plots that TC from Model 0 is much lower compared to Model 1 for the period 1995–1999. Also there is a tendency for TC to increase over time in both models. The difference in the mean values of TC between these two models is 2.7% . The difference in the pattern of TC between the models is also captured by the low (but significant) value of concordance correlation (0.521).

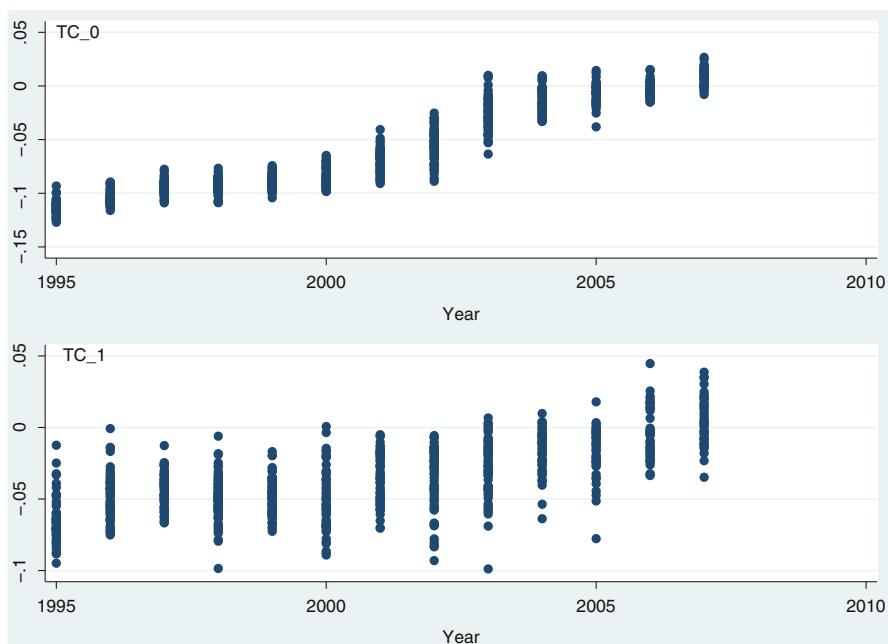


Fig. 3.2 Scatter plots of technical change in Models 0 and 1

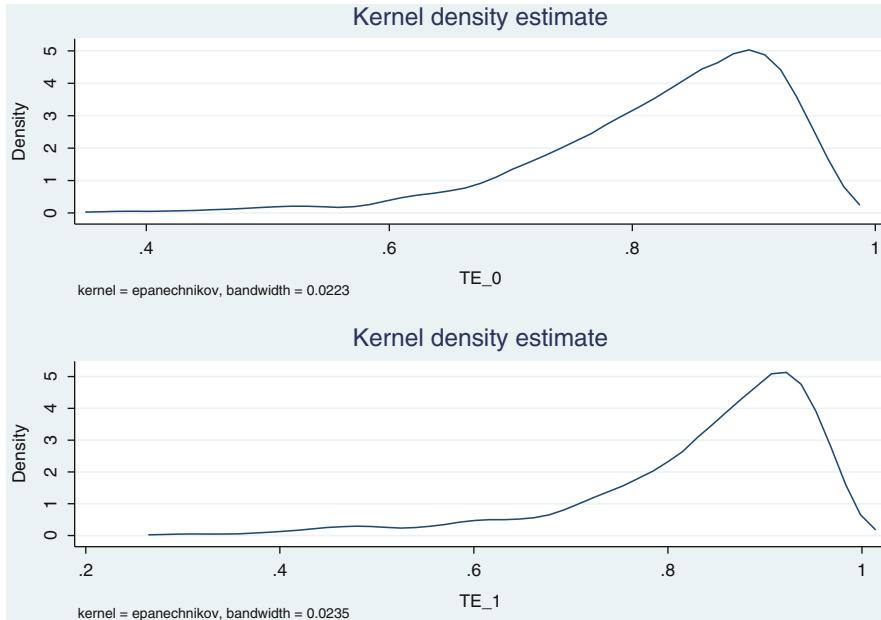


Fig. 3.3 Technical efficiency in Models 0 and 1

Next we examine TE distributions from Models 0 and 1, denoted by TE_0 and TE_1 . Density plots of TE_0 and TE_1 are reported in Fig. 3.3. Both distributions look quite similar—the distribution of TE_0 is slightly shifted to the right of the distribution of TE_1 thereby meaning the mean efficiency from Model 1 is slightly higher (83.68 % vs. 82.56 %). Again the closeness in the means and distributions of TE_0 and TE_1 based on the pooled observations might not reveal any pattern about their closeness when compared by firm and year. For this we report the scatter plots of TE_0 and TE_1 by year in Fig. 3.4. The plots reveal both cross-sectional and temporal differences in TE between the two models. It can be seen that TE_0 in the first 3 years are higher, on average, than TE_1 which tends to be much larger from 1998 onwards. There seems to be less variations (barring some outliers) in TE in the later years. A reflection of these differences can be seen in the low value of concordance correlation between TE_0 and TE_1 which is 0.094 with a standard error of 0.035.

Sometimes the interest may not be on TE scores as such but on TE changes (TEC). This is considered as catch up, i.e., the rate at which efficiency of a firm is moving to the frontier. Note that the frontier is also moving—the shift of which is captured by TC. Thus, if the catch-up rate is positive (negative) for a firm, it shows improvement (falling behind) from its current position in catching up the current year's frontier. In a regulatory case, efficiency improvement is rewarded by

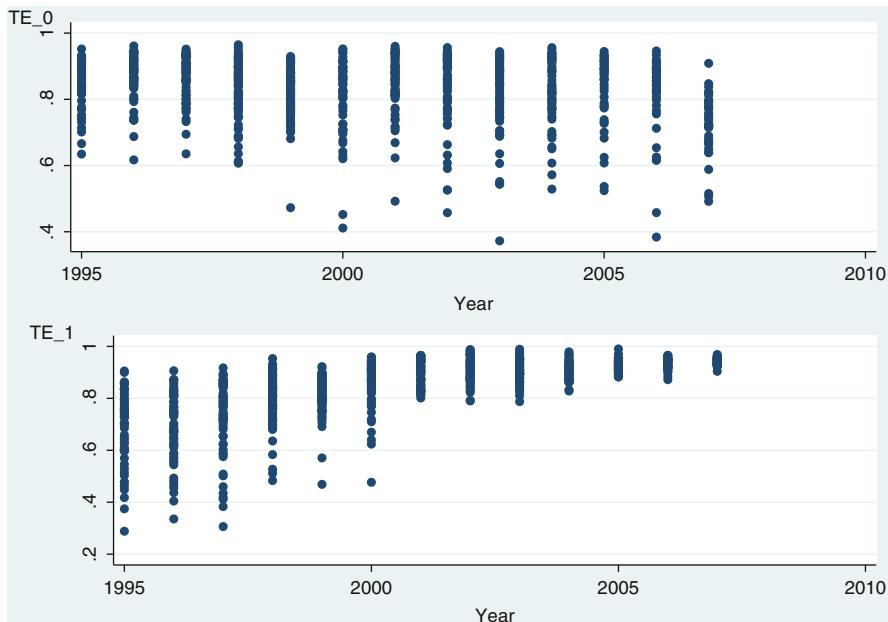


Fig. 3.4 Plots of technical efficiency in Models 0 and 1 over time

the regulators. In Fig. 3.5 we do a scatter plots of TEC^3 over time for both models (labeled as TEC_0 and TEC_1). A close look at the plots show a very different patterns of TEC_0 and TEC_1 both cross-sectionally and over time. TEC_0 is positive and do not vary much over time, that is the catch-up rate is almost constant. On the other hand, the catch-up rate in Model 1 (TEC_1) is negative (declining efficiency) which means deteriorating efficiency over time. The mean values of TEC_0 and TEC_1 are 0.9 and -3.35% with standard deviations of 0.002 and 0.023, respectively. Concordance correlation between TEC_0 and TEC_1 is 0.013 with a standard error of 0.001. Thus there is almost no agreement in the rankings of TEC between the two models.

Since both TC and TEC are measures of shift, one in the frontier and the other in efficiency, the interest is often placed on the sum of these two. This sum is known as productivity change (i.e., $PC = TC + TEC$) which in the present case can be interpreted as the rate of change in revenue over time holding everything else unchanged. The density plots of PC in Models 0 and 1 are reported in Fig. 3.6. It can be clearly seen from Fig. 3.6 that PC in these two models are very different primarily because of differences in both TC and TEC values (as discussed before).

³Note that since $TEC = \partial \ln TE / \partial t \equiv -\partial u / \partial t$, it can be viewed as a change in inefficiency with a minus sign.

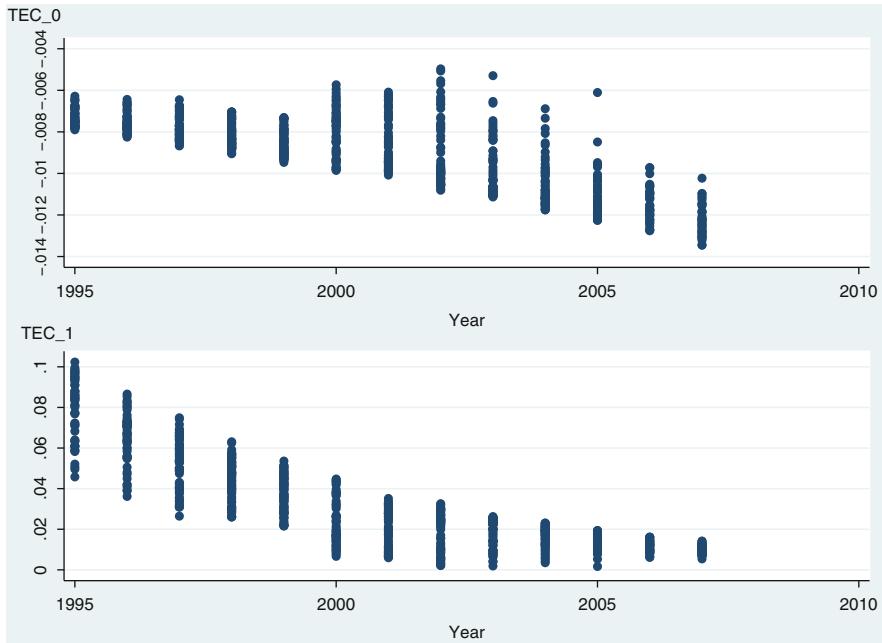


Fig. 3.5 TE Change in Models 0 and 1

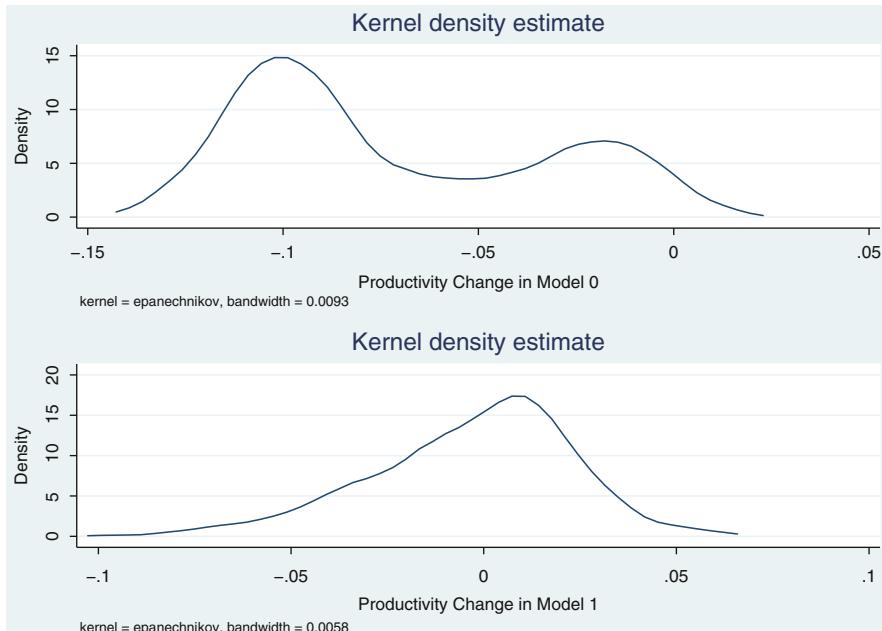


Fig. 3.6 Productivity changes in Models 0 and 1

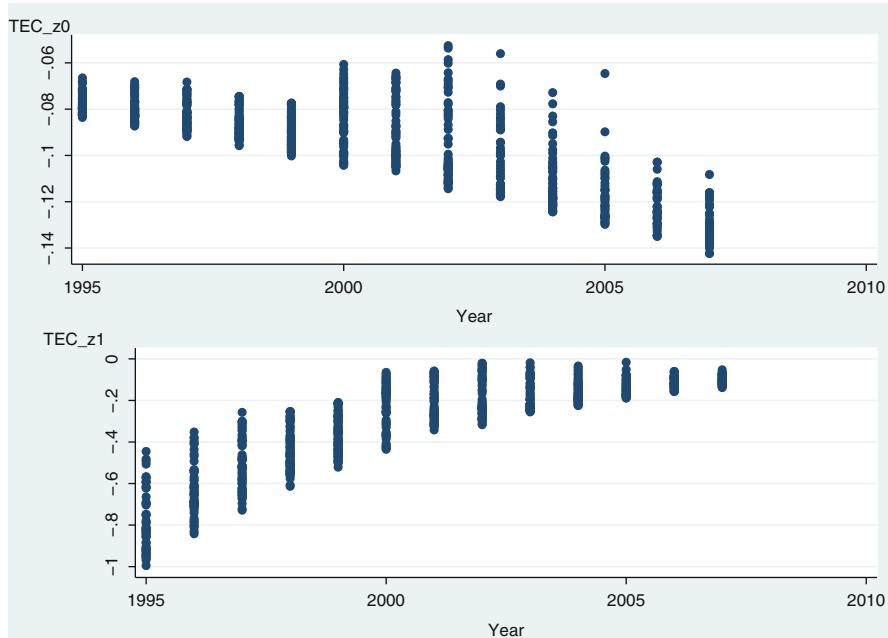


Fig. 3.7 Marginal effects of operation days in Models 0 and 1

The mean values of PC in Models 0 and 1 are -7.24 and -0.32% , respectively. The concordance rank correlation is 0.031 (standard error 0.011). This means there is almost no agreement in the rankings of PC between the two models.

In addition to the time trend variable, we also used z (log of number of operation days) to explain efficiency. The marginal effect of z on TE in Models 0 and 1 are calculated from $\partial \ln TE / \partial z$ and are reported in Fig. 3.6. These marginal effects show the percentage change in efficiency for a 1 % change in the number of operation days, and are labeled as TEC_{z0} and TEC_{z1} . The idea is to see whether number of operation days has any effect on productivity, ceteris paribus. It is clear from Fig. 3.6 that number of operation days affected negatively to productivity, and the negative effect is much higher in Model 0. In contrast Model 1 predicts positive productivity effect of number of operation days for most of the firms. To get a better picture of the productivity effect of number of operation days over time we report scatter plots of TEC_{z0} and TEC_{z1} by year in Fig. 3.7. Although both models show large variations in the productivity effect over time, it is clear that TEC_{z1} varied substantially across firms in the early years and these cross-sectional variations tend to decline over time, especially in Model 1. Further, there is a tendency of these marginal effects to decrease (increase) over time in Model 1 (Model 0). On the other hand, cross-sectional variations in TEC_{z0} are highest during 2000–2005, and it increased slightly over time. The median values of TEC_{z0} and TEC_{z1} are -0.09 and -0.26 . Thus both models show a decrease in efficiency with an increase

in number of operation days. The concordance correlation between TEC_z0 and TEC_z1 is -0.030 with a standard error of 0.003. Thus there is almost no agreement in the rankings of marginal effects of number of operation days between the two models.

3.7 Conclusion

In this paper we considered a revenue maximizing model and derived the revenue function from the transformation function where errors are parts of outputs. We used a multiplicative general error model (MGEM), in the spirit of Kumbhakar and Tsionas (2011). In our MGEM multiplicative errors are part of outputs (measurement errors). The model is further generalized to accommodate output-oriented technical inefficiency in which case the errors are viewed as the product of measurement errors and inefficiency. We used a flexible translog revenue function which in the MGEM makes the intercept and the coefficients of the linear terms random (functions of the errors associated with outputs). The errors in the revenue share functions are linear functions of the errors associated with outputs.

The vessel level data for the Norwegian whitefish fisheries for the period 1995–2007 is used to showcase the application of the MGEM. We estimate a standard (off-the-shelf) translog revenue function with output-oriented technical inefficiency and compare technical change (TC) and technical efficiency (TE) results with the MGEM revenue function which is estimated along with the revenue share equations. Although the means of TE are found to be somewhat similar, their cross-sectional patterns over time are found to be quite different. Consequently, TE change over time are also found to be quite different producing a value of concordance correlation that is around 0.5 (far from perfect agreement in their rankings). Estimates of TC are found to be negative for most of the firms in every year and are quite different across models. The estimates of productivity change are also found to be quite dissimilar with a very low measure of agreements in their rankings. Estimated productivity changes are found to negative for most of the firms in Model 0. On the contrary, these are positive for most of the firms in Model 1. Given that the specification of Model 0 (the standard revenue function) is rejected against Model 1 (the more general formulation of the revenue function), one may not take the results from Model 0 seriously.

References

- Anderson GJ, Blundell RW (1982) Estimation and hypothesis testing in dynamic singular equation systems. *Econometrica* 50:1559–1571
Asche F (2009) Adjustment cost and supply response: a dynamic revenue function. *Land Econ* 85(1):201–215

- Asche F, Bjørndal T, Gordon DV (2009) Resource rent in individual quota fisheries. *Land Econ* 85(2):279–291
- Bjørndal T, Gordon DV (1993) The opportunity cost of capital and optimal vessel size in the Norwegian fishing fleet. *Land Econ* 69:98–107
- Brown BW, Walker MB (1995) Stochastic specification in random production models of cost-minimizing firms. *J Econ* 66:175–205
- Carlson EW (1973) Cross section production functions for North Atlantic groundfish and tropical Tuna Seine fisheries. *Ocean Fishery Management: Discussions and Research, NOAA Technical report, National Marine Fisheries Service, Circ. No. 371*
- Campbell HF, Nicholl RB (1994) Can Purse Seiners Target Yellowfin Tuna? *Land Econ* 70:343–353
- Chavas J, Segerson K (1987) Stochastic specification and estimation of share systems. *J Econ* 35:337–358
- Diop H, Kazmierczak RF (1996) Technology and management in Mauritanian Cephalopod fisheries. *Mar Resour Econ* 11(2):71–84
- Guttorpsen AG, Roll KH (2011) Technical Efficiency in a Heterogeneous fishery: the case of Norwegian groundfish fisheries. *Mar Resour Econ* 26:293–308
- Jondrow J, Lovell CAK, Materov IS, Schmidt P (1982) On the estimation of technical inefficiency in the stochastic frontier production function model. *J Econ* 19:233–238
- Kirkley JE, Strand IE (1988) The technology and management of multi-species fisheries. *Appl Econ* 20:1279–1292
- Kirkley J, Squires D, Strand I Jr (1998) Characterizing managerial skill and technical efficiency in a fishery. *J Prod Anal* 9:145–160
- Kumbhakar SC, Tsionas EG (2011) Stochastic error specification in primal and dual systems. *J Appl Econ* 26:270–297
- Kumbhakar S, Asche F, Tveteras R (2013) Estimation and decomposition of inefficiency when producers maximize return to the outlay: an application to Norwegian fishing trawlers. *J Prod Anal* 40(3):307–321
- McElroy M (1987) Additive general error models for production, cost, and derived demand or share system. *J Polit Econ* 95:738–757
- Pascoe S, Coglan L, Punt AE, Dichmont C (2012) Impacts of vessel capacity reduction programmes on efficiency in fisheries. *J Agric Econ* 63:425–443
- Salvanes KG, Squires D (1995) Transferable quotas, enforcement costs and typical firms: an empirical application to the Norwegian trawler fleet. *Environ Resour Econ* 6(1):1–21
- Squires D, Kirkley J (1991) Production quotas in multiproduct fisheries. *J Environ Econ Manag* 21:109–126
- Squires D, Vestergaard N (2013) Technical change and the commons. *Rev Econ Stat* 95:1769–1787
- Thunberg EM, Bresnahan EW, Adams CM (1995) Economic analysis of technical interdependencies and the value of effort in a multi-species fishery. *Mar Resour Econ* 10(1):59–76
- Wang H-J (2002) Heteroscedasticity and non-monotonic efficiency effects of a Stochastic frontier model. *J Prod Anal* 18:241–253

Chapter 4

Production Response in the Interior of the Production Set

Loren W. Tauer

Abstract Quantile regression is used to estimate production functions at various quantiles within a dairy farm production set, and marginal products and input substitutions are derived for each of the quantile production functions. Economic relationships vary in the interior of the production set compared to the frontier of the production set, with no discernible pattern through the production set. Implication is that production response for inefficient firms in the interior of the production set may differ compared to efficient firms on the frontier of the production set as well as other inefficiency firms. The results are for a specific dairy production data set so further analysis is warranted to determine what patterns exist with other empirical production sets.

Keywords Dairy farms • Input substitution • Production efficiency • Quantile regression • Stochastic frontier analysis

4.1 Introduction

Significant efforts have been expended in recent years in measuring the inefficiency of decision making units (DMU) in production economics. The two methods commonly used have been Stochastic Frontier Analysis (SFA), where regression techniques are used to estimate a parametric frontier function from the data, and Data Envelopment Analysis (DEA), where linear programming techniques are used to envelop the data (Fried et al. 2008). Modifications of both techniques are ongoing, including parametric approaches in SFA and stochastic estimation of DEA (Kuosmanen and Johnson 2010). Much of this effort has been motivated by the desire for better measures of efficiency, or distance to the frontier for a

Paper presented at the North American Productivity Workshop VIII, Hosted by Carleton University, Ottawa, Canada, June 4–7, 2014

L.W. Tauer (✉)
Cornell University, Ithaca, NY, USA
e-mail: lwt1@cornell.edu

DMU. Less effort has been expended on economic relationships on the frontier, with little effort on the production relationships within the production set, except for potential movements of the DMU to the frontier (Serra et al. 2010). Few have attempted to investigate the economic relationship within the production set, such as measuring marginal products or input elasticities of substitution. Although these interior production points may be deemed inefficient, they none-the-less represent a production point and production response should be expected, even if still inefficient, when input usage changes.

The objective of this study then is to measure the marginal productivity and substitution of inputs within the production set using quantile regression. Quantile regression allows determining the impact of inputs on output over quantile distributions in the production set rather than only on the conditional mean as with OLS. We use data from dairy producers and find that the marginal products and the elasticities of substitution vary with movement into the production set.

4.2 Quantile Regression

Unlike ordinary least squares (OLS) and modifications of OLS, which minimize the “sum of deviations squared” around a regression line, quantile regression in contrast minimizes “least absolute deviations” around the regression line. Thus, in comparison to solving the following OLS problem: $\text{Min } \sum_{i=1}^n (y_i - \mu(x_i, \beta))^2$, quantile regression solves the following problem: $\text{Min } \sum_{i=1}^n \rho_\tau(y_i - \varphi(x_i, \beta))$, with solution by linear programming, minimizing the weighted deviations, which are weighted by ρ_τ . If ρ_τ is set equal to $1/2$ then the quantile technique produces a median rather than an average regression as in OLS. By varying the weights on the deviations, i.e. specifying various ρ_τ values ranging from 0 to 1, it is possible to generate quantile regression lines through the data set. As with OLS, the deviations in quantile regressions can be positive or negative depending upon the estimated regression line to minimize the weighted deviations. A ρ_τ value of 1 will produce a linear regression on the envelop of the data set.

Quantile regression has become popular in recent years as many realize that informative relationships can be gleaned not only from the average regression line by OLS, but also from medium and other quantile regression lines exploiting the relationships within the data set. Quantile regression has also been utilized to mute the influence of data outliers since absolute deviations are minimized rather than deviations squared as in OLS. Applications are common, but few applications have appeared in production economics (Koenker and Hallock 2001). Although their application was estimating the demand for electricity, Hendricks and Koenker in their 1992 article suggest that quantile regression may be a substitution for Stochastic Frontier Analysis. Specifically they state, “In the econometric literature on the estimation of production technologies, there has been considerable interest in

estimating so-called frontier production models that correspond closely to models for extreme quantiles of a stochastic production surface.”

There have been a few applications using quantile regression to estimate a frontier function, and then to calculate individual measures of efficiency, but the technique has not seen widespread use (Bernini et al. (2004) and Martins-Filho and Yao (2008), and others). Because quantile regression does not impose a distribution on the efficiency and error terms as does SFA, quantile regression then does not decompose the deviations into separate efficiency and random error terms as does SFA. To derive the inefficiencies of individual observations a specific quantile level must be specified as representing the frontier function. To compute efficiencies in Corrected Ordinary Least Squares (COLS), the intercept is shifted up until the estimated function intersects the highest output observation, although alternative shifts could be used. Similarly in quantile regression, one of the upper quantiles must be selected as representing the frontier. Any quantile regression can be selected, but in keeping with the practice of hypothesis testing at the 95 % level, often the 95 % quantile is established as the frontier function (Behr 2010).

The applications of quantile regression to estimating efficiency have focused on the estimation of efficiency rather than the estimation of marginal products of inputs, or input substitution. To the author’s knowledge, none of the quantile regression production applications have estimated input substitution elasticities. Behr (2010) for instance, used quantile regression to estimate bank efficiency from a Cobb-Douglas specification, and only reported and plotted production elasticities. There are a number of reasons for this omission besides the primary interest in efficiency estimation. Given the quantile selected to represent efficient production, there may be very few efficient observations. Thus movement along the frontier may have little empirical support. Also, because quantile regression estimates quantile rather than conditional mean relationships, if the error distribution is homoscedastic, then over the various quantile estimates only the intercept and not the slope coefficients would be expected to change. The implication is that marginal products and elasticities of substitution across the quantiles would not change.

In this paper I estimate input substitution elasticities and marginal products not only for the frontier production function, but also for various quantile production functions in the interior of the production set. I find variability in these estimates with the dairy production data set used.

4.3 Marginal Products and Elasticities

The Cobb-Douglas production function for three inputs is specified as:

$$\ln y = \alpha_0 + \sum_{i=1}^3 \alpha_i \ln x_i \text{ where } y \text{ is the output and } x_1, x_2 \text{ and } x_3 \text{ are the inputs.}$$

The α_i term represents the elasticity of production for input i, which is the marginal product of input i divided by the average product of input i.

In contrast, the Translog function for one output and three inputs is specified as:

$$\ln y = \alpha_0 + \sum_{i=1}^3 \alpha_i \ln x_i + \frac{1}{2} \sum_{i=1}^3 \sum_{j=1}^3 \beta_{ij} \ln x_i \ln x_j$$

This is a more flexible form than the Cobb-Douglas and allows non-unitary elasticities of substitution between inputs.

Marginal products from the Translog function are then a function of all inputs.

$$MP_j = \frac{\partial y}{\partial x_j} = \frac{y}{x_j} \frac{\partial \ln(y)}{\partial \ln(x_j)} = \frac{y}{x_j} \left(\alpha_j + \sum_i \beta_{ij} \ln(x_i) \right)$$

The Allen Elasticities (AES) from the Translog production function are derived as:

$$\sigma_{ij} = \frac{\sum_k f_k}{x_i x_j} \frac{x_k}{F} F_{ij}$$

Where F is the bordered Hessian matrix of the Translog production function and F_{ij} is the ij cofactor of that bordered Hessian. Although it is customary to compute marginal products and elasticities at the geometric means of the data, we compute these at each data observation to arrive at marginal products and elasticities throughout the production set.

4.4 Data

Data were from the New York Dairy Farm Business Summary for the calendar year 2011 (Knoblauch et al. 2012). These are data from 211 farms that were identified as owner operated dairy farms where at least 85 % of the gross income originated from milk sales. Because these are full time dairy farms, by-products of milk production, such as cull dairy cows and excess calves sold, were added to output rather than treated as separate outputs, which would have required estimating a distance function. Some farmers also sell some crops if they experience a good crop production year resulting in more dairy feed than necessary to feed the herd, and those crops sales are also added to milk receipts.

Three inputs are defined—capital flow, labor, and material. Individual components of these three inputs are listed in Table 4.1. Capital flow includes the service flow from the cattle stock, machinery, and real estate. Although land is often modeled as a separate input in agriculture production, the value of land is not estimated by farmers separately from the value of real estate in the data survey. Dairy operations in New York have significant investments in dairy barns and milking facilities which can exceed the value of the farmland. Labor includes the imputed value of operators' labor (as determined by each operator), the value of family labor contributed, and wages paid. Material primarily includes purchased feed and inputs

Table 4.1 Dairy farm business summary data variables 2011

Variable	Components	Average (2011 \$1000)	S.D. (2011 \$1000)
Milk (output)	Milk sales minus milk marketing expenses, by-product sales (cull cows, calves sold, excess crops, custom work payments, direct government payments)	2922	3119
Capital flow (input)	Replacement livestock, expansion livestock, breeding expenses, cattle supplies, veterinary, machinery rent, machinery repair, machinery depreciation, rent, building repair, real estate tax, property insurance, real estate depreciation, 4 % interest on (average livestock inventory, average machinery investment, and average real estate)	683	676
Labor (input)	Value of operators and family labor, wages	443	475
Materials (input)	Purchased feed and roughage, fertilizer, seeds, pesticides, fuel and utilities, bedding, milking supplies, bovine somatotropin (BST), misc. expenses	1341	1435

to produce crops, both of which are fed to cows to produce milk. All measures are accrual reflecting what was actually used in the business during the calendar year 2011. Plots of the log of output individually against the logs of capital, labor, and material are shown in Fig. 4.1. These relationships are increasing as expected, but output variation heterogeneity appears to exist over various capital and labor levels, which is not as pronounced in the use of material. The implication is that quantile regressions may vary in slope coefficients on the input variables, especially capital and labor, which indeed occurs.

4.5 Results

Estimation and computations are done in R using the packages “quantreg” for quantile regression (Koenker 2013), “frontier” for SFA (Coelli and Henningsen 2013) and the native OLS “lm” routine in R for linear regression. Computation of the marginal products and input elasticities were facilitated using the “micEcon” package in R (Henningsen 2014). Although the Cobb-Douglas is restricted to unitary elasticities of substitution between inputs, the Cobb-Douglas provides good first order estimates of the marginal products (elasticities) and is the functional form often used when estimating a production function. Thus the Cobb-Douglas is first estimated by OLS. The empirical estimates are shown in Table 4.2. The overall fit of the model is high with an adjusted R squared value of 0.96 with highly statistical significant t-statistics on the estimated coefficients. The three production elasticities sum to 1.058 which implies increasing returns to scale.

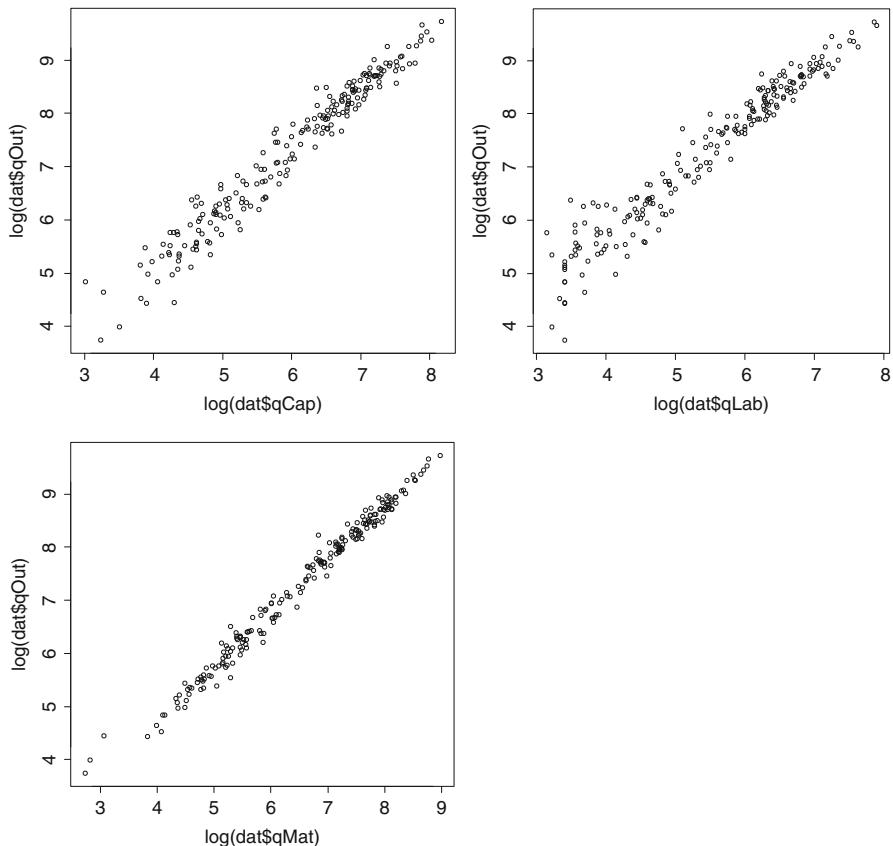


Fig. 4.1 Log output plotted against log of capital, labor, and material

Table 4.2 Cobb-Douglas production function estimated by OLS

Coefficients	Estimate	Std. error	t-Value
(Intercept)	0.664	0.045	14.93
Log (qCap)	0.268	0.029	9.23
Log (qLab)	0.125	0.026	4.74
Log (qMat)	0.665	0.027	24.90
Residual standard error:	0.1196	on 207 DF	
Multiple R-squared:	0.9923	adjusted R-squared:	0.9922
F-statistic:	8907	on 3 and 207 DF,	p-value: < 2.2e-16

Quantile regression of the Cobb-Douglas production function with tau values ranging from 0.1 to 0.9 in decile increments summarized in Fig. 4.2 shows the estimated coefficient elasticities vary over the quantiles. In Fig. 4.2 the solid horizontal lines represent the OLS estimates as reported in Table 4.2, and the dashed horizontal lines represent the confidence intervals. The non-horizontal lines

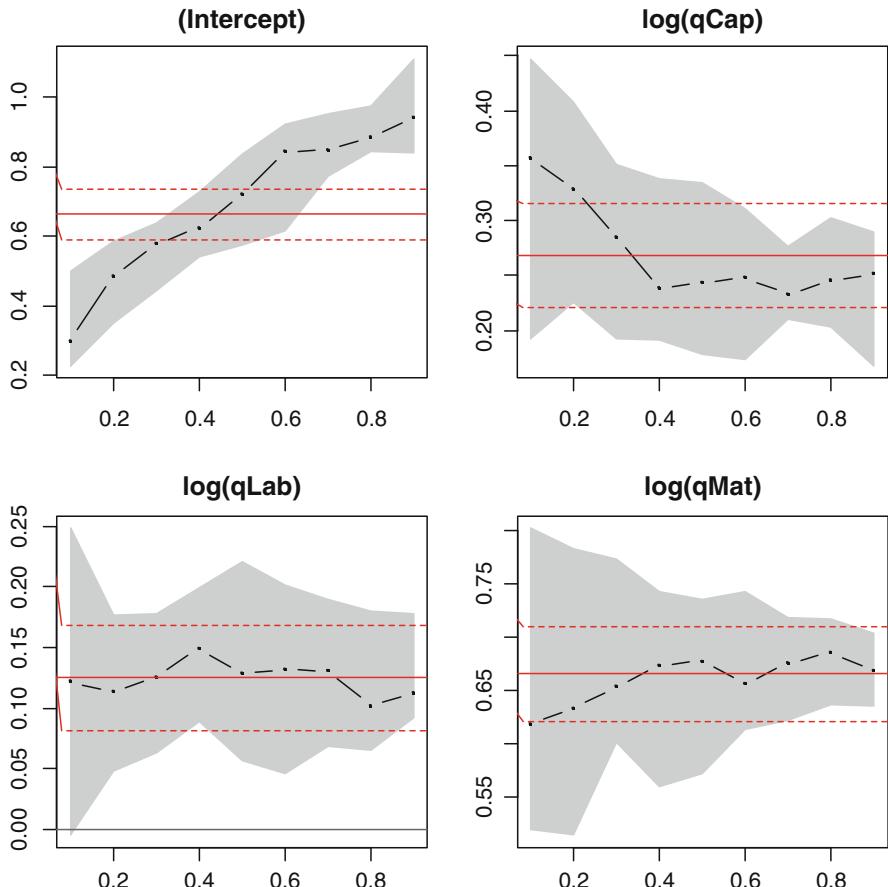


Fig. 4.2 Cobb-Douglas elasticities of inputs from quantile regressions from 0.1 to 0.9

represent the quantile coefficient estimates over the quantiles, while the shaded line represents the confidence interval of those estimates. The quantile point estimates are for the most part contained in the confidence interval of the OLS estimates except for the intercept estimates.

The Translog production functions were estimated also using actual observations rather than deviations from geometric means because marginal products and elasticities were desired for each netput value rather than only at their geometric means. The Translog estimates by OLS are shown in Table 4.3. The overall fit of the model is still high with an adjusted R squared value of 0.99, although the statistical significance of the individual linear coefficients is reduced by spreading the response impact over the quadratic terms. A Likelihood Ratio Test of the Cobb-Douglas with the Translog, where the null was restricting the quadratic terms to be zero, produced a Chi-Square value of 45.42, clearly supporting the Translog over the Cobb-Douglas.

Table 4.3 Translog production estimated by OLS

Coefficient	Estimate	Std. error	t-Value
(Intercept)	-0.015	0.202	-0.08
Log (qCap)	0.140	0.142	0.99
Log (qLab)	0.110	0.130	0.84
Log (qMat)	1.003	0.116	8.63
(0.5 * log (qCap)^2)	0.274	0.128	2.15
(0.5 * log (qLab)^2)	0.185	0.113	1.63
(0.5 * log (qMat)^2)	0.375	0.081	4.64
(Log (qCap) * log (qLab))	0.068	0.091	0.75
(Log (qCap) * log (qMat))	-0.287	0.090	-3.20
(Log (qLab) * log (qMat))	-0.208	0.074	-2.80
Residual standard error: 0.1129 on 201 DF			
Multiple R-squared: 0.9933, adjusted R-squared: 0.993			
F-statistic: 3331 on 9 and 201 DF, p-value: < 2.2e-16			

Table 4.4 Stochastic frontier translog production function (efficiency as half normal)

Coefficient	Estimate	Std. error	z-Value
(Intercept)	0.115	0.195	0.59
Log (qCap)	0.101	0.135	0.75
Log (qLab)	0.108	0.124	0.87
Log (qMat)	1.034	0.110	9.38
I (0.5 * log (qCap)^2)	0.269	0.120	2.24
I (0.5 * log (qLab)^2)	0.162	0.109	1.48
I (0.5 * log (qMat)^2)	0.372	0.081	4.61
I (log (qCap) * log (qLab))	0.085	0.088	0.96
I (log (qCap) * log (qMat))	-0.291	0.087	-3.36
I (log (qLab) * log (qMat))	-0.202	0.072	-2.80
SigmaSq	0.021	0.004	4.97
Gamma	0.658	0.143	4.59
Log likelihood value: 167.6473			

Behr (2010) in a quantile estimate of bank efficiency also found the Translog functional form statistically increased the explanatory power according to an F-test, but he elected to use the Cobb-Douglas given its R^2 value close to 1.

The estimated Translog Stochastic Frontier Analysis function estimated with a half normal for inefficiency is shown in Table 4.4. The estimates of the SFA coefficients change slightly from the OLS coefficient estimates, but the statistical significance of variables do not change. The material coefficient and the quadratic terms involving material are statistically significant in both the OLS and the SFA estimates, as are the quadratic terms on capital and labor (weakly).

The quantile estimates of the Translog at tau values of 1.00, 0.95, 0.90, 0.80, 0.70, 0.50 and 0.25 are summarized in Table 4.5. High tau values were selected given the interest in the production response near the frontier of the production set, but the median regression of 0.50 is also reported as well as the low quantile value of 0.25.

Table 4.5 Quantile regression translog production estimates (t-statistic in parentheses)

Coefficients	Quantile regression tau values						
	<u>1.00</u>	<u>0.95</u>	<u>0.90</u>	<u>0.80</u>	<u>0.70</u>	<u>0.50</u>	<u>0.25</u>
(Intercept)	0.313 (0.71)	0.037 (0.10)	0.453 (1.08)	0.615 (1.87)	0.215 (0.63)	0.057 (0.17)	-0.738 (-1.87)
Log (qCap)	0.677 (1.58)	0.435 (1.07)	0.206 (0.80)	0.081 (0.31)	0.125 (0.52)	0.076 (0.26)	0.397 (1.14)
Log (qLab)	-0.175 (-0.42)	-0.273 (-0.87)	0.036 (0.14)	0.128 (0.60)	0.204 (0.91)	0.191 (0.88)	0.110 (0.51)
Log (qMat)	0.734 (2.15)	1.115 (2.88)	0.931 (2.81)	0.900 (4.17)	0.899 (3.91)	0.973 (4.57)	0.956 (4.04)
I (0.5 * log (qCap)^2)	0.751 (2.29)	0.310 (0.77)	0.076 (0.24)	0.067 (0.32)	0.067 (0.34)	0.327 (1.65)	0.290 (1.14)
I (0.5 * log (qLab)^2)	-0.182 (-0.57)	0.150 (0.61)	0.084 (0.39)	-0.037 (-0.19)	0.128 (0.68)	0.133 (0.87)	0.190 (1.00)
I (0.5 * log (qMat)^2)	0.803 (3.54)	0.457 (1.47)	0.137 (0.53)	0.194 (1.10)	0.355 (2.12)	0.338 (2.82)	0.447 (3.39)
I (log (qCap) * log (qLab))	0.224 (0.93)	0.119 (0.63)	0.057 (0.32)	0.146 (0.94)	0.167 (1.24)	0.030 (0.26)	0.078 (0.65)
I (log (qCap) * log (qMat))	-0.924 (-3.87)	-0.424 (-1.21)	-0.116 (-0.41)	-0.158 (-1.08)	-0.187 (-1.26)	-0.296 (-2.48)	-0.342 (-2.11)
I (log (qLab) * log (qMat))	0.006 (0.03)	-0.163 (0.94)	-0.095 (-0.49)	-0.103 (-0.65)	-0.264 (-1.81)	-0.142 (-1.20)	-0.223 (-1.40)

Shown are the coefficient estimates and below each coefficient the corresponding t-statistic. The t-statics are from bootstrapped estimates of the standard errors as implemented in “quantreg” with the null hypotheses of zero coefficient values.

The marginal products using the Translog functional form with the various models are summarized in Table 4.6. Marginal products for each observation is computed and then averaged. Shown are marginal product from quantile regression estimates at tau values of 1.00, 0.95, 0.90, 0.80, 0.70, 0.50 and 0.25, as well as SFA estimates with a half normal for efficiency and OLS estimates. Marginal products in most cases are slightly greater than one, reflecting that more than a dollar of total output revenue is produced per dollar of input, although a few are less than one, and there is significant variation across the models, especially for capital and labor; less so for material. This reflects the output heterogeneity over capital and labor illustrated in Fig. 4.1. The implication is that the response along the estimated production function differs depending upon position within the production set. Note that these marginal products are from incremental changes in inputs along an estimated production function which may be in the interior of the production set. The explicit assumption is that although the marginal product may be measured at an inefficient point, that marginal product does not incorporate any change in efficiency unless efficiency change occurs due to a movement along the estimated production function.

Table 4.6 Marginal products of capital, labor, and material on New York dairy farms estimated from a translog production function by quantile regressions, stochastic frontier analysis and OLS

Technique	Capital	Labor	Material
Quantile, tau = 1.00	1.484	1.755	1.151
Quantile, tau = 0.95	0.642	1.390	1.660
Quantile, tau = 0.90	0.917	1.506	1.502
Quantile, tau = 0.80	1.050	0.890	1.550
Quantile, tau = 0.70	0.891	1.087	1.482
Quantile, tau = 0.50	0.945	1.020	1.329
Quantile, tau = 0.25	1.083	0.838	1.178
SFA	1.105	1.149	1.422
OLS	1.016	1.030	1.299

Table 4.7 Allen elasticities of substitution between capital, labor, and materials on New York dairy farms estimated from a translog production function by quantile regressions, stochastic frontier analysis, and OLS

Technique	Capital-labor	Capital-material	Labor-material
Quantile, tau = 1.00	0.648	-0.444	0.293
Quantile, tau = 0.95	9.502	-0.167	-1.272
Quantile, tau = 0.90	-3.232	5.725	4.939
Quantile, tau = 0.80	17.729	-10.864	-11.915
Quantile, tau = 0.70	3.562	-0.647	-1.383
Quantile, tau = 0.50	10.108	-2.304	1.395
Quantile, tau = 0.25	-0.335	-0.318	-1.636
SFA	-20.280	-2.324	0.791
OLS	14.682	1.147	-3.315

Allen elasticities of substitution reported in Table 4.7 vary immensely across models, and any two inputs may be either substitutes or complements depending upon the estimated function. This instability of elasticities may be a reflection of the low statistical significant of some of the individual coefficients of the Translog function, which changes with movement through the production set. Note, however, that even with the excellent fit of the OLS Cobb-Douglas with individual coefficients with high t-values, these coefficients still vary at various quantile estimates as shown in Fig. 4.2. If producers' substitution possibilities are subject to the quantile production regressions, then the substitutability of inputs very much depends upon where the netput vector resides in the production set. However, changing inputs may also change efficiency, which would alter the elasticity of substitution.

These farms are very efficient as shown by the kernel density plots in Fig. 4.3 of the efficiency scores estimated by the SFA function and by the observations below the 0.95 tau and the 0.50 tau value quantile regressions. The SFA was estimated with a half normal specified for efficiencies. The SFA efficiencies are more narrowly distributed and that distribution lies to the right of the tau = 0.95 quantile efficiency distribution. The tau = 0.50 regression measures farms as higher efficient and thus that distribution lies to the right of the tau = 0.95 efficiency distribution. A lower tau value for the quantile regression moves the regression line lower into the production set and increases the measured quantile efficiency of an observation.

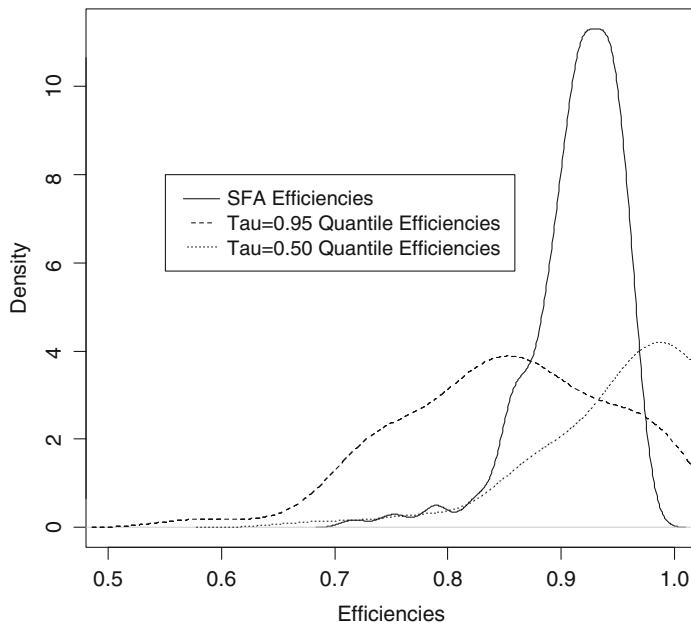


Fig. 4.3 Efficiency scores for stochastic frontier and quantile ($\tau = 0.95$ and 0.50) regressions

A plot of the individual farm SFA efficiencies against the quantile regression ($\tau = 0.95$) farm efficiencies are shown in Fig. 4.4. That shows a positive relationship with a correlation of 0.89. A 45° line imposed on Fig. 4.4 shows that the SFA technique measures efficiency higher for most of the farms except when the farm efficiency is greater than about 0.95, which is the τ value of the quantile regression line used to measure quantile regression efficiency.

4.6 Conclusion

We find that marginal products and elasticities of substitution are different in the interior of the production set at various quantile regressions compared to these relationships either on the frontier of the production set or from an average regression. Quantile regression is used to estimate regression equations of various quantiles within the interior of the production set populated by data observations. A Cobb-Douglas function produced excellent fit, but a Translog was statistically superior, although the t-statistics on the Translog function decreased due to the underlying multicollinearity of that functional form across inputs. Empirical results were derived from a data set of dairy producers. Additional efforts would be useful to determine if these results are unique or more universal across various products and regions.

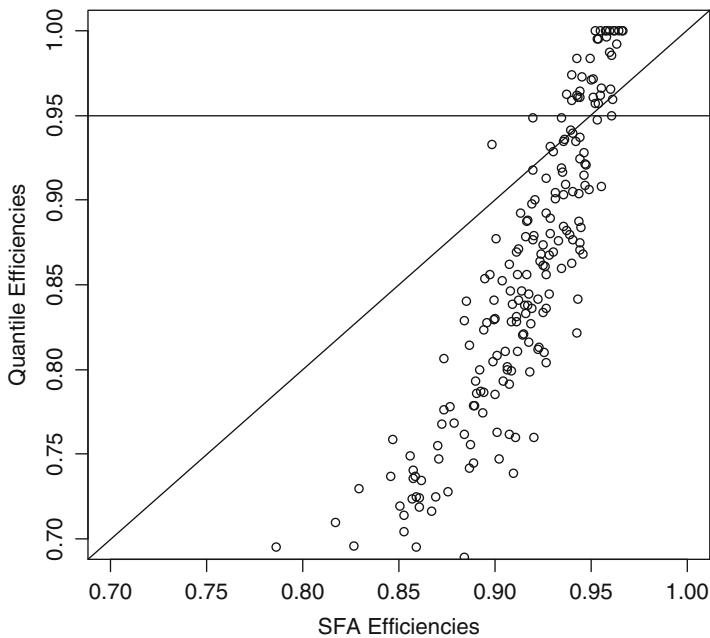


Fig. 4.4 Plot of SFA efficiencies against quantile regression efficiencies ($\tau = 0.95$) (correlation = 0.89)

References

- Behr A (2010) Quantile regression for robust bank efficiency score estimation. *Eur J Oper Res* 200:568–581
- Bernini C, Freo M, Gardini A (2004) Quantile estimation of frontier production function. *Empir Econ* 29:373–381
- Coelli T, Henningsen A (2013) Frontier: stochastic frontier analysis. R package version 1.1-0. <http://CRAN.R-Project.org/package=frontier>. Accessed 16 Apr 2015
- Fried HO, Knox Lovell CA, Schmidt SS (eds) (2008) The measurement of productive efficiency and productivity growth. Oxford University Press, New York
- Hendricks W, Koenker R (1992) Hierarchical spline models for conditional quantiles and the demand for electricity. *J Am Stat Assoc* 87:58–68
- Henningsen A (2014) micEcon: microeconomic analysis and modelling. R package version 0.6-12. <http://CRAN.R-project.org/package=micEcon>. Accessed 16 Apr 2015
- Knoblauch WA, Putnam LD, Karszes J, Overton R, Dymond C (2012) Business summary New York state 2011, R.B. 2012-01, Charles H. Dyson School of Applied Economics and Management, Cornell University
- Koenker R. (2013) Quantreg: quantile regression. R package version 5.05. <http://CRAN.R-project.org/package=quantreg>. Accessed 16 Apr 2015
- Koenker R, Hallock KF (2001) Quantile regression. *J Econ Perspect* 15:143–156
- Kuosmanen T, Johnson A (2010) Data envelopment analysis as nonparametric least squares regression. *Oper Res* 58:149–160
- Martins-Filho C, Yao F (2008) A smooth nonparametric conditional quantile frontier estimator. *J Econ* 143:317–333
- Serra T, Lansink AO, Stefanou S (2010) Measurement of dynamic efficiency: a directional distance function parametric approach. *Am J Agri Econ* 93:752–763

Chapter 5

Spillover Effects of Public Capital Stock Using Spatial Frontier Analyses: A First Look at the Data

Jaepil Han, Deockhyun Ryu, and Robin C. Sickles

Abstract This paper aims to investigate spillover effects of public capital stock in a production function model that accounts for spatial dependencies. Although there are a number of studies that estimate the output elasticity of public capital stock, they suffer from a failure to refine the output elasticity of public capital stock as well as to account for spillover effects of the public capital stock on the production efficiency when such spatial dependencies exist. For this purpose we employ a spatial autoregressive stochastic frontier model and analyze estimates with a time-varying spatial weights matrix. Using data for 21 OECD countries from 1960 to 2001, we found that spillover effects can be an important factor explaining variations in technical inefficiency across countries as well as discrepancies among various levels of output elasticity of public capital stock in traditional production function approaches.

Keywords Public capital • Spillover effects • Stochastic frontier model • Spatial panel model • Time-varying spatial weights

5.1 Introduction

Public capital includes many types of goods which are used to produce final goods and services for consumers. The infrastructures such as highways, streets, roads, and public educational buildings take the largest components of public capital and also electric, gas and water supply facilities, administration, police, military service, hospital facilities, and many other forms of goods and services are included in public capital. In the United States, the real government gross fixed capital formation is

J. Han (✉) • R.C. Sickles
Department of Economics, Rice University, Houston, TX, USA
e-mail: jh24@rice.edu; rsickles@rice.edu

D. Ryu
Department of Economics, Chung-Ang University, Seoul, S. Korea
e-mail: dhryu@cau.ac.kr

between 3 and 4 % of real GDP and about 20 % of the real gross fixed capital formation in the private sector. Based on the estimates of Kamps (2006), the public capital stock is sizable, taking about 55 % of real GDP and more than 20 % of private capital stock on average of the 22 OECD countries during the sample periods: 1960–2001. This implies that ignoring public capital can be problematic when analyzing productivity and efficiency. Delorme et al. (1999) also argue that public infrastructure reduces the technical inefficiency of private-sector production.

The economic impact of public investment has been received great attention in numerous studies for the last few decades. Many studies have tried to estimate the output elasticity of public capital and there has been a debate on the effects of public capital on output. The output elasticity estimates of public capital from various models and samples range from 0.1 to 0.4. Specifically, the earlier studies report relatively large elasticity estimates (Aschauer 1989; Munnell 1990) while they are considered implausibly high by subsequent studies (Tatom 1991; Holtz-Eakin 1994). Recently, Bom and Lighthart (2013) estimate an average output elasticity of public capital to be around 0.15 using meta-analysis with 68 studies for the 1983–2008 periods. Even though the magnitudes of the effects of public capital stock are far from consensus, there is little doubt on the positive sign and statistical significance (Pereira and Andraz 2013). There have been several explanations for the disagreement on the magnitude of the effects of public capital. Major explanations are related to econometric issues, such as omitted variable bias and possibility of spurious regressions (Tatom 1991), or ignorance of cross sectional dependency, which results in inefficient and biased estimates and invalid inference when it actually exists. Another explanation is the possible existence of spillover effects and the different levels of samples of studies. Pereira and Andraz (2013) point out that the spillover effects captured by aggregate level studies can give a clue for the disagreement on the effects of public capital. The issue of the possible existence of spillovers of public capital has received relatively little attention, even though some studies examine the geographic spillover effects of public investment using local data (Holtz-Eakin and Schwartz 1995; Pereira and Roca-Sagales 2003). Recently spatial econometric methods are extensively used in regional science studies to assess spatial spillovers.

After the pioneer spatial autoregressive model by Cliff and Ord (1973), the spatial econometric models are extended to panel data models and estimation techniques have been developed (Anselin 1988; Kapoor et al. 2007; Lee and Yu 2010b). Spatial econometrics consists of econometric techniques dealing with interactions of economic units in space. These interactions can be of geographical or economic characteristics. The spatial weights matrices tend to be time invariant because mostly spatial weights are based on geographic concepts such as border sharing characteristics or centroid distances, which are not change over time. However, we can consider the economic/socioeconomic distances or demographic characteristics, which might change over time. Spatial dependency and the heterogeneity of the spatial dependence structure can influence on the productivity or efficiency of the economic units, but the standard stochastic frontier models do not take spatial interactions into account. Recent studies try to incorporate the spatial dependency

into a general specification frontier model (Pavlyuk 2012; Glass et al. 2013a, b 2014a, b; Adetutu et al. 2015).

In this paper, we mainly concern how to estimate the effects of public capital stock on output separating out the direct effects and the indirect effects. The direct effects include the feedback effects which pass through neighboring regions and back to the region itself. The indirect effects are interpreted as spillover effects. We consider the possible existence of spatial dependence by incorporating the spatially correlated terms with a variety of spatial dependence structures. Also, we measure the technical efficiencies of sample countries. We expect to improve the estimation of the technical efficiency by adding public capital as a factor input and controlling the possible cross sectional dependence. To this end we investigate the quasi-maximum likelihood (QML) estimation of Spatial Autoregressive Stochastic Frontier Model. Finally, we apply the model to a dataset from 21 OECD countries under the setting of time-varying spatial weights matrix. We found significant and sizable output elasticity of public capital, and significant spillover effects, and also we estimated the relative technical efficiency scores of each models.

The paper continues with the following structure. Section 5.2 introduces the standard spatial models and presents the associate frontier model we are interested in. Also, we discuss on the direct, indirect, and total effects of the inputs on output and connect the interpretation to the spillover effects. In Sect. 5.3, we modify the quasi-maximum likelihood estimation provided by Lee and Yu (2012) for the efficiency analysis. In Sect. 5.4, we apply the model to a dataset from 21 OECD countries. Finally, Sect. 5.5 concludes the paper.

5.2 Spatial Autoregressive Stochastic Frontier Model

We begin with a non-spatial production function. The production function is of the form:

$$y_{it} = \beta_0 + X_{it}\beta + \varepsilon_{it}, \quad i = 1, \dots, N; \quad t = 1, \dots, T \quad (5.1)$$

where i indexes cross-section of economic units and t indexes time periods. y_{it} is output of the i th unit at time t , whereas X_{it} is a $(1 \times K)$ input vector of the i th unit at time t . β is the $(K \times 1)$ parameter vector to be estimated, and ε_{it} is an *i.i.d.* disturbance for i and t with zero mean and variance σ_ε^2 .

In general, three types of spatial interaction effects can be given on the non-spatial production function. The first are endogenous interaction effects, which are explaining dependence between the dependent variable, y , of each unit. The second are exogenous interaction effects, which are explaining dependence between the dependent variable of a specific unit, y , and the independent variable of another unit, X . The last type is the interaction effects among the error terms. The generic form of the model with all types of spatial effects can be written as:

$$\begin{aligned} Y &= \rho WY + \alpha \iota_n + X\beta + WX\theta + \varepsilon, \\ \varepsilon &= \lambda W\varepsilon + u, \end{aligned} \quad (5.2)$$

where W is a known non-negative element of the $(N \times N)$ spatial weights matrix, WY denotes the endogenous interaction effects, WX the exogenous interaction effects, and $W\varepsilon$ the interaction effects among the disturbance term.¹

Among the interaction effects, we are interested in the endogenous interaction effects. Hence we impose the restrictions of $\theta = 0$ and $\lambda = 0$. Then the model of interest is:

$$y_{it} = \rho \sum_{j=1}^N w_{ij} y_{jt} + \beta_0 + X_{it}\beta + \varepsilon_{it}, \quad (5.3)$$

where ε_{it} is an *i.i.d.* disturbance for i and t with zero mean and variance σ_ε^2 . The model is called Spatial Autoregressive Model; hereafter SAR. This is the Cliff-Ord type production function, suggested by Cliff and Ord (1981).

For the efficiency analysis, we need to transform the Cliff-Ord type production function (5.3) to an associated frontier model by introducing non-negative random variable u_i which represents the technical inefficiency of unit i . We assume u_i to be time-invariant following Schmidt and Sickles (1984). ε_{it} is divided into two parts: u_i , non-negative random variable associated with technical inefficiency, and v_{it} , a systematic random error.

$$y_{it} = \rho \sum_{j=1}^N w_{ij} y_{jt} + \beta_0 + X_{it}\beta - u_i + v_{it}, \quad (5.4)$$

Define $\alpha_i \equiv \beta_0 - u_i$, then the model becomes

$$y_{it} = \rho \sum_{j=1}^N w_{ij} y_{jt} + \alpha_i + X_{it}\beta + v_{it}. \quad (5.5)$$

A relative inefficiency (or efficiency) measure accounts for the output of each unit compared to the output that could be produced by a fully-efficient unit. Because the most efficient unit has the largest α_i , the relative inefficiency measure can be derived by defining u_i^* as the distance between $\max(\hat{\alpha}_i)$ and $\hat{\alpha}_i$,²

$$u_i^* \equiv \max(\hat{\alpha}_i) - \hat{\alpha}_i. \quad (5.6)$$

¹Elhorst (2014) named this equation as the general nesting spatial (GNS) model.

²Since output is in logarithms, relative technical efficiency is defined as $\hat{r}_i \equiv \exp(-u_i^*)$:

We can write the stacked form of (5.5) as follows:

$$Y_t = \rho W Y_t + X_t \beta + \alpha + V_t, \quad t = 1, \dots, T, \quad (5.7)$$

where $Y_t = (y_{1t}, y_{2t}, \dots, y_{Nt})'$ and $V_t = (v_{1t}, v_{2t}, \dots, v_{Nt})'$ are $(N \times 1)$ column vectors, and v_{it} 's are *i.i.d.* across i and t with zero mean and variance σ_v^2 . The X_t is an $(N \times K)$ matrix of non-stochastic regressors, and α is an $(N \times 1)$ column vector of individual effects. The spatial weights matrix W_t is a non-stochastic and row-normalized, and time varying $(N \times N)$ matrix.

The spatial weights matrix, W , is taken to be exogenous and captures cross-section dependence among observations. The spatial weights matrix is mostly specified to be time invariant because it is usually based on the characteristics that are hardly changing over time such as geographic distance or border sharing feature. However, it can be formed from other concepts that are time varying such as economic/socioeconomic distances or demographic characteristics. We incorporate time varying characteristics of the spatial weights matrix as follows:

$$Y_t = \rho W_t Y_t + X_t \beta + \alpha + V_t. \quad (5.8)$$

The reduced form of (5.8) is

$$Y_t = (I_N - \rho W_t)^{-1} (X_t \beta + \alpha + V_t). \quad (5.9)$$

One advantage of incorporating the spatial dependence is that it captures the direct and indirect effects separately. The estimation of the non-spatial model only returns the estimates of total effects of factor inputs without identifying the direct and indirect effects. When a spatial lag term is introduced in the model, the direct effects of an explanatory variable, X_k , are the diagonal elements of $(I_N - \rho W_t)^{-1} \beta_k$, while the indirect effects are the off-diagonal elements of the matrix. The pre-multiplied matrix can be decomposed as:

$$(I_N - \rho W_t)^{-1} = I_N + \rho W_t + \rho^2 W_t^2 + \rho^3 W_t^3 + \dots. \quad (5.10)$$

Hence the direct effect will be greater than or equal to β_k . The first two matrix terms of the right hand side of (5.10) represent a direct effect of a change in X_k only and an indirect effect of a change in X_k only, respectively. The rest terms represent higher order direct and indirect effects, which include the feedback effects of other units. To obtain a single direct effect and indirect effect for an explanatory variable in the model, LeSage and Pace (2009) suggest reporting the average of the direct effects and the average of the indirect effects.

To test whether spatial spillovers exist or not, the estimated indirect effects should be used, not the estimate of ρ . This is because the indirect effects are derived from the multiplication of $(I_N - \rho W_t)^{-1}$ and β_k , and the variation of the indirect effects depends on the variation of all coefficient estimates. In other words, even though each coefficient is estimated to be significant, this does not mean that

the mean indirect effect is significant, and vice versa. LeSage and Pace (2009) suggest simulating the distribution of the direct and indirect effects. The simulation procedure is basically in two steps: (1) computing the mean value over D draws of direct/indirect effects for the approximation of the overall effects and (2) obtaining t-statistics by dividing the sample mean by the corresponding standard deviation.

5.3 QML Estimation of Spatial Autoregressive Model with Time-Varying Spatial Weights Matrices

Spatial models can be estimated by maximum likelihood, quasi-maximum likelihood, instrumental variables, generalized method of moments, or by Bayesian Markov Chain Monte Carlo methods. One advantage of QML is that it does not rely on the assumption of normality of the disturbances. For the standard panel data model with fixed effects, one can estimate jointly the common parameters of interest and fixed effects by the maximum likelihood estimation. However, it is well-known that the MLE of the variance is inconsistent when T is finite. Similar consequences are found for the spatial panel data model with fixed effects. To avoid the incidental parameter problem, we can use a data transformation, which is a demeaning procedure of each variable. Lee and Yu (2010b) provide asymptotic properties of quasi-maximum likelihood estimators for spatial dynamic panel data with both time and individual fixed effects. Elhorst (2014) summarizes how to estimate spatial autoregressive model with fixed effects when spatial weight matrix is time invariant. In this section, we derive quasi-maximum likelihood estimator for spatial autoregressive model with fixed effects when spatial weight matrix is time dependent.

Let $\theta = (\beta, \rho, \sigma_v^2)'$. The log-likelihood function of the model (5.9) is

$$\text{Log}L_{N,T}(\theta, \alpha) = -\frac{NT}{2} \ln(2\pi\sigma_v^2) + \sum_{t=1}^T \ln |I_N - \rho W_t| - \frac{1}{2\sigma_v^2} \sum_{t=1}^T V'_t(\theta) V_t(\theta), \quad (5.11)$$

where $V_t(\theta) = (I_N - \rho W_t) Y_t - X_t \beta - \alpha$.

We can find α_i which maximize the log-likelihood function as follows:

$$\alpha_i = \frac{1}{T} \sum_{t=1}^T \left(y_{it} - \rho \sum_{j=1}^N w_{ij}^t y_{jt} - X_{it} \beta \right), \quad i = 1, \dots, N. \quad (5.12)$$

By substituting (5.12) into (5.11), we concentrate out α in (5.11) and get the concentrated log-likelihood with α concentrated out as follows:

$$\text{Log}L_{N,T}(\theta) = -\frac{NT}{2} \ln(2\pi\sigma_v^2) + \sum_{t=1}^T \ln |I_N - \rho W_t| - \frac{1}{2\sigma_v^2} \sum_{t=1}^T \tilde{V}'_t(\theta) \tilde{V}_t(\theta),$$

where $\tilde{V}_t(\theta) = \tilde{Y}_t - \rho \widetilde{W_t Y_t} - \tilde{X}_t \beta$, $\tilde{Y}_t = Y_t - \frac{1}{T} \sum_{t=1}^T Y_t$, $\widetilde{W_t Y_t} = W_t Y_t - \frac{1}{T} \sum_{t=1}^T W_t Y_t$, and $\tilde{X}_t = X_t - \frac{1}{T} \sum_{t=1}^T X_t$.

It can be shown that the estimates for β and σ_v^2 can be expressed in functions of the autoregressive coefficient ρ :

$$\beta(\rho) = (\tilde{X}'_t \tilde{X}_t)^{-1} \tilde{X}'_t [\tilde{Y}_t - \rho \widetilde{W_t Y_t}], \quad (5.14)$$

$$\sigma_v^2(\rho) = \frac{1}{N(T-1)} (\tilde{Y}_t - \rho \widetilde{W_t Y_t} - \tilde{X}_t \beta)' (\tilde{Y}_t - \rho \widetilde{W_t Y_t} - \tilde{X}_t \beta) \quad (5.15)$$

Substitution of (5.14) and (5.15) into (5.13) returns a concentrated likelihood function, which contains only one unknown parameter, ρ :

$$\text{Log}L_{N,T}(\rho) = C - \frac{NT}{2} \ln [(\tilde{e}_0 - \rho \tilde{e}_1)' (\tilde{e}_0 - \rho \tilde{e}_1)] + \sum_{t=1}^T \ln |I_N - \rho W_t|, \quad (5.16)$$

where C is a constant not relying on ρ , \tilde{e}_0 and \tilde{e}_1 are the residuals corresponding to the regression of \tilde{Y}_t and $\widetilde{W_t Y_t}$ on \tilde{X}_t , respectively.³

Finally, by maximizing (5.16), we can obtain a solution of ρ . Even though a closed-form solution of ρ does not exist, the numerical solution is unique because the concentrated log-likelihood function is concave in ρ . Once we have the estimator of ρ , we can compute the estimator of β and σ_v^2 by using $\hat{\rho}$ for ρ in (5.14) and (5.15).

For the asymptotic properties of the QML estimators, we need a number of regularity conditions for the spatial weight matrix, W_t , which are detailed in Han et al. (2015), who also provide formulas for the asymptotic variance matrix of the parameters.

5.4 An Empirical Application

In this section, we apply the SASF model to a dataset from 21 OECD Countries⁴ for the period 1960–2001. For the analysis, we construct the spatial weights

³The estimator of σ_v^2 is bias-corrected as suggested by Lee and Yu (2010a) to avoid possible bias caused by the demeaning procedure.

⁴Australia, Austria, Belgium, Canada, Denmark, Finland, France, Greece, Iceland, Ireland, Italy, Japan, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom, and United States.

matrices using both geographic distance and economic distance. Also, for each cross sectional relations, we try contiguity approach and distance decaying approach with and without threshold for the weights matrix construction. Finally, we allow the weights matrices change over time.

5.4.1 Specifications of the Spatial Weights Matrices

Prior to the estimation, we need to specify the spatial dependence structure between observations. The approaches to construct the spatial weight matrices often used in practice are roughly categorized into two groups: weights based on a contiguity or distance. Typically geographic relations are used for specifying the dependence, but we can use economic relations. As Lee and Yu (2012) point out, when spatial weights matrix is constructed from those characteristics in a panel or dynamic setting, the characteristics might change over time. In this paper, we constructed the bilateral trade volume to exploit the economic relations between countries. The distance concepts we use in this section are as follows:

Yearly Bilateral Trade: D_E^t

$$d_{ij}^t = ex_{ij}^t + im_{ij}^t.$$

The economic distance concept has advantages over typical geographic distance or time-invariant distance in that it provides more affluent information. However, most of the elements of the weights matrices are non-zero, so if the dimension gets larger, the numerical calculation is burdensome with those weights matrix. To avoid such a problem, one can generate distance weights with thresholds such as k nearest neighbors.⁵ To satisfy the regularity conditions for the asymptotic analysis, the distance concepts are row-normalized before the usage. The normalized weights matrices are notated as W instead of D , which represents the non-normalized weights matrices.

5.4.2 Data Description

For the estimation of the typical production function, we need four key variables: real GDP, labor participation, private capital stock, and public capital stock. The data are obtained by manipulating variables extracted from Penn World Table Version 7.1 (Heston et al. 2012). For the estimation of private capital stock, we use the perpetual inventory method after obtaining real aggregate investment. In using the perpetual inventory method, we assume that the real capital stock in 1960 is depreciated real

⁵For more examples, see Han et al. (2015).

Table 5.1 Summary statistics

	Mean	Std. dev	Min	Max
Real GDP (millions, US\$)	875.8	1598.5	3.9	11,265.5
Labor participation (thousands)	16,448	26,188	79	144,746
Private capital (millions, US\$)	1941.3	3434.6	7.5	24,753.4
Public capital (millions, US\$)	528.0	1004.8	2.1	5705.4

aggregate investment in 1959 and the fixed depreciation rate of 6 % for all countries for all periods. Major challenge of empirical study on public capital productivity is shortage of public capital data. Kamps (2006) provides the public capital estimates of 22 OECD countries from 1960 to 2001. Unfortunately, the study is outdated so the estimates are not compatible with the up-to-date Penn World Table dataset and also the estimates are in national currencies, which is not suitable for international comparison.⁶ Hence we calculated the public capital stock using the public capital ratio to real GDP in his study and the real GDP from Penn World Table. The summary statistics of the four key variables are shown in Table 5.1. To construct the spatial weight matrices, we collected the bilateral trade data from NBER-UN Trade Data 1962–2000 (Feenstra et al. 2005). The dataset is based on reports by the importing country assuming that they are more accurate than reports by the exporting countries.⁷

5.4.3 Empirical Findings

For benchmarks, we estimated the non-spatial production function (5.1) by Pooled OLS and Fixed Effects technique. The estimation results are displayed in the first two columns of Table 5.2. The coefficients of the three factor inputs are significantly different from zero and have the expected signs. The non-spatial fixed effects model does not capture direct and indirect effects separately. Even though the estimated results of SASF are displayed on the last column of Table 5.2, we are not able to compare the coefficient estimates in the non-spatial model and their counterparts of the models under spatial setting one on one. This is because the coefficient estimates in the spatial autoregressive stochastic frontier model are no longer interpreted as output elasticities of the factor inputs because of the existence of $(I_N - \rho W_t)^{-1}$ in (5.9). For appropriate comparisons, we need to obtain the mean direct effects, mean indirect effects and total effects as we discussed in Sect. 5.2.

⁶It is hard to find appropriate price indexes and PPP exchange rates for all countries for the periods.

⁷For a limitation of the dataset, it has several data points that indicate a country imports from itself. This happens because the several series related to international trade are aggregated by simply summed up. For the regularity conditions, we ignored the ‘self-trade’ data points.

Table 5.2 Estimation results

	Pooled OLS	Fixed effects	SASF
Intercept	4.830	–	–
	(6.491)	–	–
Log (L)	0.441	0.489	0.197
	(5.857)	(15.922)	(5.811)
Log (Kp)	0.287	0.240	0.167
	(4.549)	(22.279)	(10.903)
Log (Kg)	0.266	0.302	0.253
	(2.826)	(15.191)	(12.550)

Note: The numbers in parentheses are t-stats

In Table 5.3, we present the direct and indirect effects estimates. The direct effects are computed by averaging the diagonal elements of $(I_N - \rho W_t)^{-1} \beta_k$, and the indirect effects are computed by averaging the row sums of the off-diagonal elements of $(I_N - \rho W_t)^{-1} \beta_k$ with 1000 parameter combinations draws. We constructed the spatial weights matrix using the yearly bilateral trade volume hence we can compute the output elasticities over time. We are almost not able to observe the variations of the output elasticities across the period. This is because the main trade partners of the most countries do not vary much even though each country prefers different countries for their partners. The total effects of the factor inputs range from 0.293 to 0.300 for labor, from 0.251 to 0.253 for private capital, from 0.380 to 0.384 for public capital. The indirect effects are around 33 % of the total effects for all factor inputs across the period. We find that the total effects of labor are smaller than the results of the time invariant weights specifications, while the total effects of private capital and public capital are greater than the corresponding total effects. Especially, the differences in the total effects of labor are mainly due to the difference in the direct effects, while the difference in the total effects of private and public capital are due to the difference in the indirect effects.

Let us turn our attention to the results of the technical efficiency analyses. With the estimation results, we obtained the efficiency scores and rankings of countries. Table 5.4 displays the relative efficiency scores estimates and the rankings. In the non-spatial fixed effects model, Iceland appears to be the most efficient country, while Japan is the least efficient country. However, we regard the estimation results as biased because we found the spatial lag model is more appropriate from the Lagrange multiplier tests, hence the technical efficiency estimates are also biased. The results from the spatial model show relatively low efficiency scores than those from non-spatial fixed effects model. Moreover, the efficiency scores change dramatically when we incorporate the spatial autoregressive term. Considering the economic relations between countries, United States is the most efficient in terms of productivity, while Iceland is the least efficient country. This represents that the ignoring cross-sectional dependence can mislead the efficiency analyses. From the change between the non-spatial Fixed Effects model and SASF, we conclude that

Table 5.3 Direct and indirect effects with time-varying economic weights matrix

	Year	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	
	Log (L)																						
Total	0.295	0.296	0.297	0.296	0.294	0.293	0.294	0.296	0.296	0.295	0.297	0.297	0.297	0.297	0.297	0.295	0.295	0.298	0.296	0.3	0.296		
(6.36)	(6.19)	(6.45)	(6.39)	(6.5)	(6.22)	(6.25)	(6.28)	(6.23)	(6.26)	(6.28)	(6.26)	(6.24)	(6.21)	(6.24)	(6.24)	(6.24)	(6.24)	(6.24)	(6.24)	(6.24)	(6.24)	(6.47)	
Direct	0.198	0.199	0.2	0.199	0.198	0.197	0.197	0.197	0.199	0.199	0.199	0.199	0.2	0.2	0.2	0.2	0.2	0.199	0.198	0.201	0.199	0.202	0.199
(6.03)	(5.74)	(5.97)	(5.97)	(5.98)	(5.78)	(5.82)	(5.79)	(5.79)	(5.79)	(5.79)	(5.79)	(5.79)	(5.79)	(5.79)	(5.79)	(5.79)	(5.79)	(5.79)	(5.79)	(5.79)	(5.79)	(5.79)	
Indirect	0.097	0.096	0.097	0.097	0.096	0.095	0.097	0.097	0.097	0.097	0.097	0.097	0.097	0.097	0.097	0.097	0.096	0.096	0.097	0.097	0.098	0.096	
Log (Kp)																							
Total	0.251	0.252	0.252	0.251	0.252	0.252	0.251	0.251	0.252	0.252	0.251	0.251	0.252	0.252	0.253	0.253	0.252	0.252	0.252	0.251	0.251	0.252	
(15.18)	(14.88)	(15.26)	(15.08)	(15.56)	(14.52)	(15.06)	(14.64)	(14.82)	(15.05)	(14.67)	(15.08)	(15.40)	(15.61)	(15.11)	(15.1)	(14.98)	(14.87)	(14.67)	(14.87)	(14.87)	(15.00)		
Direct	0.169	0.17	0.17	0.169	0.17	0.169	0.17	0.169	0.17	0.169	0.17	0.169	0.17	0.17	0.17	0.17	0.17	0.169	0.17	0.169	0.169	0.17	
(11.13)	(11.02)	(11.43)	(11.25)	(11.71)	(10.64)	(11.18)	(10.75)	(11.06)	(11.37)	(10.74)	(11.32)	(11.42)	(11.00)	(11.06)	(11.07)	(11.00)	(11.22)	(11.14)	(11.25)	(11.14)	(11.25)		
Indirect	0.083	0.082	0.082	0.082	0.082	0.083	0.083	0.083	0.082	0.082	0.082	0.082	0.082	0.082	0.082	0.082	0.083	0.082	0.082	0.082	0.082	0.082	
Log (Kg)																							
Total	0.354	0.382	0.381	0.384	0.383	0.383	0.384	0.384	0.382	0.381	0.382	0.382	0.381	0.382	0.382	0.381	0.382	0.383	0.381	0.384	0.383	0.382	
(11.79)	(11.50)	(12.15)	(11.81)	(11.93)	(11.42)	(11.95)	(11.51)	(11.22)	(11.80)	(11.65)	(11.51)	(11.83)	(11.87)	(11.77)	(11.41)	(11.46)	(11.47)	(11.47)	(11.42)	(11.51)	(11.65)		
Direct	0.257	0.257	0.256	0.257	0.257	0.257	0.257	0.257	0.256	0.257	0.257	0.257	0.257	0.257	0.257	0.257	0.256	0.257	0.257	0.257	0.257	0.257	
(12.69)	(12.39)	(12.79)	(12.49)	(12.24)	(13.07)	(12.90)	(12.54)	(12.25)	(12.58)	(12.24)	(12.58)	(12.24)	(12.85)	(12.59)	(12.73)	(12.63)	(12.54)	(12.22)	(12.45)	(12.58)			
Indirect	0.127	0.125	0.125	0.126	0.125	0.127	0.127	0.127	0.126	0.125	0.126	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.126	0.125		
Log (Kd)																							
Total	(6.50)	(6.45)	(6.65)	(6.62)	(6.76)	(6.32)	(6.65)	(6.29)	(6.11)	(6.58)	(6.39)	(6.51)	(6.53)	(6.66)	(6.40)	(6.31)	(6.38)	(6.46)	(6.53)	(6.53)	(6.60)		
Direct	0.198	0.198	0.2	0.199	0.197	0.198	0.2	0.201	0.199	0.199	0.199	0.199	0.199	0.199	0.2	0.199	0.198	0.199	0.199	0.2	0.199		
(5.80)	(5.71)	(5.86)	(5.96)	(5.81)	(5.96)	(5.77)	(6.04)	(5.75)	(5.74)	(5.93)	(5.94)	(5.72)	(5.87)	(5.93)	(5.88)	(5.88)	(5.91)	(5.78)	(6.09)	(5.86)	(5.87)		
Indirect	0.096	0.096	0.097	0.097	0.096	0.096	0.096	0.097	0.098	0.096	0.096	0.097	0.097	0.096	0.097	0.096	0.097	0.097	0.097	0.097	0.097		
Log (Lp)																							
Total	0.294	0.295	0.297	0.296	0.295	0.293	0.295	0.298	0.299	0.295	0.295	0.295	0.296	0.295	0.295	0.296	0.295	0.295	0.295	0.297	0.296		
(6.39)	(6.32)	(6.38)	(6.36)	(6.22)	(6.47)	(6.16)	(6.48)	(6.12)	(6.16)	(6.48)	(6.14)	(6.47)	(6.14)	(6.17)	(6.36)	(6.41)	(6.41)	(6.34)	(6.48)	(6.32)	(6.41)		
Direct	0.198	0.198	0.2	0.199	0.197	0.198	0.2	0.201	0.199	0.199	0.199	0.199	0.199	0.199	0.2	0.199	0.198	0.199	0.199	0.199	0.2		
(5.92)	(5.82)	(5.81)	(5.89)	(5.61)	(5.96)	(5.69)	(5.66)	(5.51)	(5.56)	(5.51)	(5.56)	(5.51)	(5.56)	(5.56)	(5.37)	(5.62)	(5.62)	(5.62)	(5.68)	(5.68)	(5.74)		
Indirect	0.096	0.096	0.097	0.097	0.096	0.096	0.096	0.097	0.098	0.096	0.096	0.097	0.097	0.096	0.097	0.097	0.096	0.097	0.097	0.097	0.097		
Log (Kg)																							
Total	0.252	0.251	0.252	0.251	0.253	0.252	0.251	0.252	0.252	0.252	0.252	0.251	0.251	0.252	0.251	0.251	0.252	0.252	0.253	0.252	0.251		
(14.72)	(14.36)	(15.57)	(14.50)	(14.73)	(15.15)	(15.52)	(14.24)	(15.11)	(15.25)	(15.23)	(14.48)	(15.01)	(14.33)	(14.33)	(15.51)	(15.51)	(15.51)	(15.26)	(16.04)	(15.24)	(15.23)		
Direct	0.17	0.169	0.17	0.169	0.17	0.169	0.17	0.169	0.17	0.169	0.17	0.169	0.17	0.169	0.17	0.169	0.17	0.169	0.17	0.17	0.169		
(10.95)	(10.47)	(11.47)	(10.83)	(11.84)	(11.37)	(11.66)	(10.81)	(11.40)	(11.24)	(11.32)	(10.83)	(11.08)	(11.08)	(11.08)	(11.39)	(11.39)	(11.39)	(11.88)	(11.88)	(11.88)	(11.10)		
Indirect	0.082	0.082	0.082	0.082	0.083	0.082	0.082	0.083	0.083	0.082	0.083	0.082	0.082	0.082	0.082	0.083	0.082	0.083	0.082	0.083	0.083		
Log (Kd)																							
Total	0.382	0.383	0.382	0.383	0.384	0.382	0.383	0.382	0.383	0.382	0.383	0.382	0.383	0.382	0.383	0.383	0.383	0.383	0.383	0.384	0.384		
(11.23)	(10.91)	(11.63)	(11.35)	(11.71)	(12.15)	(12.12)	(11.02)	(11.96)	(11.80)	(13.51)	(11.79)	(11.55)	(10.98)	(12.12)	(12.01)	(11.63)	(12.51)	(11.94)	(11.68)				
Direct	0.257	0.257	0.256	0.257	0.258	0.256	0.257	0.257	0.256	0.256	0.257	0.256	0.257	0.257	0.257	0.257	0.257	0.256	0.256	0.256	0.257		
(12.17)	(12.21)	(12.78)	(12.46)	(13.11)	(12.79)	(11.85)	(12.82)	(12.73)	(12.87)	(12.41)	(12.78)	(12.34)	(11.56)	(11.49)	(12.80)	(13.05)	(12.81)	(12.70)	(12.77)				
Indirect	0.125	0.126	0.125	0.126	0.126	0.126	0.126	0.126	0.126	0.126	0.126	0.126	0.126	0.126	0.126	0.126	0.126	0.126	0.126	0.126	0.127		
Log (Kp)																							
Total	(6.38)	(6.00)	(6.42)	(6.40)	(6.43)	(6.62)	(6.79)	(6.37)	(6.78)	(6.50)	(6.53)	(6.37)	(6.43)	(6.61)	(6.26)	(6.71)	(6.67)	(6.85)	(6.73)	(6.36)	(6.36)		

Table 5.4 Efficiency scores estimates

	Non-spatial FE		SASF	
	Eff. score (%)	Ranking	Eff. score (%)	Ranking
Australia	77.3	8	25.2	11
Austria	77.7	13	23.4	13
Belgium	86.8	4	28.4	9
Canada	87.3	3	23.9	12
Denmark	72.7	12	22.2	15
Finland	69.7	16	20.4	16
France	73.0	11	52.5	3
Greece	67.8	19	20.3	17
Iceland	100.0	1	7.7	21
Ireland	68.2	18	12.1	20
Italy	70.9	15	47.6	4
Japan	51.6	21	36.5	6
Netherlands	76.1	9	32.0	7
New Zealand	69.4	17	12.4	19
Norway	83.2	6	23.1	14
Portugal	66.8	20	19.9	18
Spain	76.1	10	37.4	5
Sweden	83.8	5	31.7	8
Switzerland	87.6	2	26.6	10
United Kingdom	71.0	14	53.9	2
United States	78.3	7	100.0	1

smaller economies get relatively less benefits from other countries, while larger economies get more benefits from other countries. For the periods, the average GDP of Iceland is the smallest among the countries.

For the distributional comparisons of the models, we plot the kernel densities of the efficiency scores in Fig. 5.1. We found the distribution from the non-spatial fixed effects model located slightly to the right side of the density plots of SASF. Finally, in Fig. 5.1b and c, we compared the efficiency scores distributions when we include the public capital stock as a factor input and exclude the public capital stock. We observe the efficiency scores distributions are shifted to the right when we add public capital as a factor input, which implies that the inclusion of public capital stock variable helps to explain some of variations in production.

5.5 Conclusions

In this paper, we investigate spillover effects of public capital stock in a production frontier model that accounts for cross sectional dependency among countries. We estimate the output elasticity of public capital stock as well as labor and private

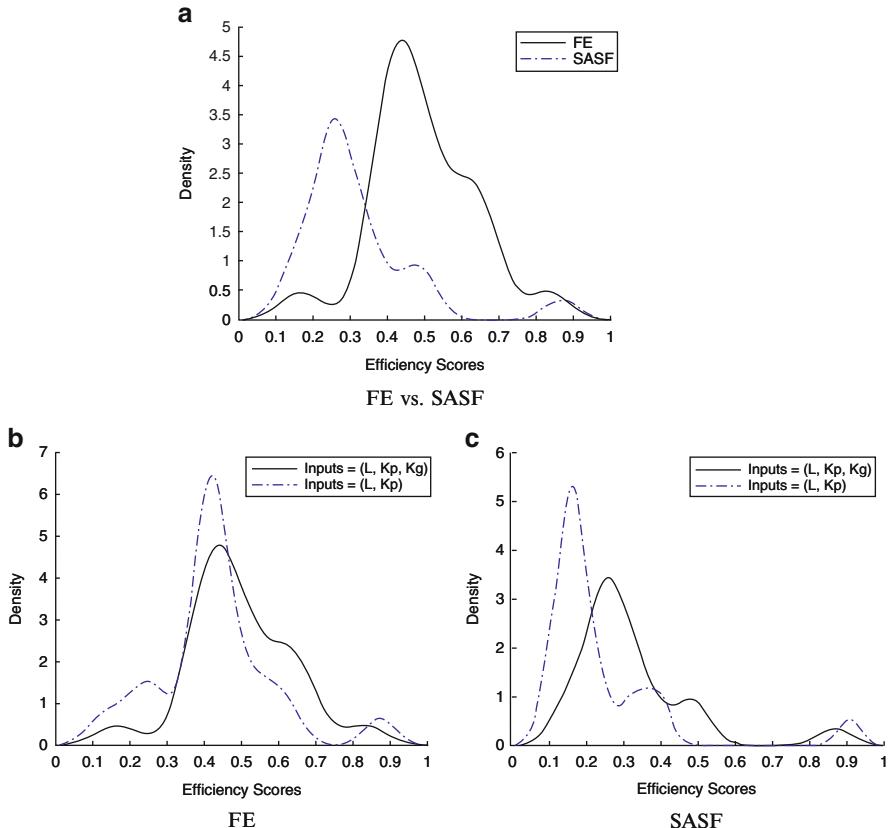


Fig. 5.1 Kernel densities of efficiency scores (a) FE vs. SASF. (b) FE. (c) SASF

capital stock in a Cliff-Ord type production frontier function. Especially, we separate out the direct and indirect effects of factor inputs and interpret the indirect effects as the spillovers. Moreover, the spatial autoregressive stochastic frontier model allows us to gain the relative efficiencies of countries as well. The spatial weights matrix is the essential characteristic of the approach and we exploit the economic interactions to construct our spatial weights matrix. Specifically, we allow for the spatial weights matrix to vary across time, which is reasonable for economic/socioeconomic spatial weights.

We estimate the output elasticities of factor inputs at international level by the empirical application to the data for 21 OECD countries from 1960 to 2001. We found relatively large total output elasticities of public capital stock and observe that the indirect effects comprise around 35 % of the total elasticities, which supports the argument of Pereira and Andraz (2013). Finally, we measure the relative technical efficiencies of countries. Concerning the spatial dependency,

SASF model is expected to give the less biased relative efficiency estimates than the non-spatial counterpart, but more delicate comparison between efficiency estimates from different models should be analyzed by the future works.

References

- Adetutu M, Glass AJ, Kenjegalieva K, Sickles R (2015) The effects of efficiency and TFP growth on pollution in Europe: a multistage spatial analysis. *J Prod Anal* 43(3):307–326
- Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Dordrecht
- Aschauer DA (1989) Is public expenditure productive? *J Monet Econ* 23:177–200
- Bom PR, Lighthart JE (2013) What have we learned from three decades of research on the productivity of public capital? *J Econ Surv* 28(5):889–916
- Cliff A, Ord J (1973) Spatial autocorrelation. Pion, London
- Cliff A, Ord J (1981) Spatial processes, models and applications. Pion, London
- Delorme CD, Thompson HG, Warren RS (1999) Public infrastructure and private productivity: a stochastic-frontier approach. *J Macroecon* 21:563–576
- Elhorst JP (2014) Spatial econometrics: from cross-sectional data to spatial panels. Springer, Heidelberg
- Feenstra RC, Lipsey RE, Deng H, Ma AC, Mo H (2005) World trade flows: 1962–2000. Working paper 11040. National Bureau of Economic Research, Cambridge
- Glass AJ, Kenjegalieva K, Sickles R (2013a) A spatial autoregressive production frontier model for panel data: with an application to European countries, mimeo. Measuring Economic Performance: Theory and Practice: International Workshop and the Launch of the Efficiency and Productivity Research Interest Group, Loughborough University, Loughborough
- Glass AJ, Kenjegalieva K, Sickles R (2013b) How efficiently do U.S. cities manage roadway congestion? *J Prod Anal* 40(3):407–428
- Glass AJ, Kenjegalieva K, Sickles R (2014a) A spatial autoregressive production frontier model for panel data with asymmetric efficiency spillovers. Working paper, Rice University
- Glass AJ, Kenjegalieva K, Sickles R (2014b) Estimating efficiency spillovers with state level evidence for manufacturing in the US. *Econ Lett* 123(2):154–159
- Han J, Ryu D, Sickles R (2015) How to measure spillover effects of public capital stock: a spatial autoregressive stochastic frontier model. Working paper, Rice University
- Heston A, Summers R, Aten B (2012) Penn world table version 7.1. Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania
- Holtz-Eakin D (1994) Public-sector capital and the productivity puzzle. *Rev Econ Stat* 76:12–21
- Holtz-Eakin D, Schwartz AE (1995) Spatial productivity spillovers from public infrastructure: evidence from state highways. *Int Tax Public Financ* 2:459–468
- Kamps C (2006) New estimates of government net capital stocks for 22 OECD countries, 1960–2001. *IMF Staff Pap* 53:120–150
- Kapoor M, Kelejian HH, Prucha IR (2007) Panel data models with spatially correlated error components. *J Econ* 140:97–130
- Lee L, Yu J (2010a) Estimation of spatial autoregressive panel data models with fixed effects. *J Econ* 154:165–185
- Lee L, Yu J (2010b) A spatial dynamic panel data model with both time and individual fixed effects. *Economic Theory* 26:564–597
- Lee LF, Yu J (2012) QML estimation of spatial dynamic panel data models with time varying spatial weights matrices. *Spat Econ Anal* 7(1):31–74
- LeSage J, Pace RK (2009) Introduction to spatial econometrics. Chapman & Hall, Boca Raton
- Munnell AH (1990) Why has productivity declined? Productivity and public investment. *N Engl Econ Rev*, January/February 3–22

- Pavlyuk D (2012) Maximum likelihood estimator for spatial stochastic frontier models. In: 12th International Conference “Reliability and Statistics in Transportation and Communication”, Riga, Latvia, 11–19
- Pereira AM, Andraz JM (2013) On the economic effects of public infrastructure investment: a survey of the international evidence. *J Econ Dev* 38(4):1–37
- Pereira AM, Roca-Sagales O (2003) Spillover effects of public capital formation: evidence from the Spanish regions. *J Urban Econ* 53:238–256
- Schmidt P, Sickles RC (1984) Production frontiers and panel data. *J Bus Econ Stat* 2(4):367–374
- Tatom JA (1991) Public capital and private sector performance. *Fed Res Bank St Louis Rev* 73:3–15

Chapter 6

Dynamic Technical Efficiency

Lynda Khalaf and Charles J. Saunders

Abstract Dynamic panels with non-Gaussian errors suffer from the incidental parameter bias. Simulations show that an indirect inference estimation approach provides bias correction for the model and distribution parameters. The indirect confidence set inference method is size correct and exhibits good coverage properties even for asymmetric confidence regions. Bank cost data are examined under the proposed dynamic technical efficiency framework with evidence that an MLE approach could provide misleading implications.

Keywords Indirect inference • Dynamic models • Confidence set • Monte Carlo

6.1 Introduction

Several methods have been proposed to correct for the bias including, but not limited to, GMM, bias-correction terms, indirect inference, and X-differencing.

Although there are relatively few restrictions on the error distributions, in general Gaussian errors are assumed. The dynamic panel approach allows for less restrictive error distributions, including non-Gaussian like the skew-normal distribution. Relatively few studies take into account the possibility of dynamics in the panel structure or errors, see Ahn and Sickles (2000), Desli et al (2003) and Tsionas (2006).

The first contribution of this paper demonstrates that the inclusion of dynamics into a stochastic frontier model can lead to biased parameter estimates. Maximum likelihood estimators that are designed to take into account the incidental parameter bias are few. The implications of this bias on a MLE approach is examined via simulations, which demonstrates that the parameter bias is large. The MLE parameter bias has little effect on the estimation of the distribution parameter related to skewness.

L. Khalaf (✉) • C.J. Saunders
Department of Economics, Carleton University, 1125 Colonel By Drive,
Ottawa, ON, Canada K1S 5B6
e-mail: Lynda.Khalaf@carleton.ca; Charles.Saunders@carleton.ca

The second contribution is the introduction of the indirect inference estimation (IIE) approach for bias correction. The approach extends that of Gouriéroux et al (2010) to incorporate the additional parameters by allowing for skew-Normal errors. This extension leads to bias correction of the lagged dependent parameter and relatively unbiased skewness parameter estimates. The confidence intervals for the parameters are constructed with the indirect confidence set inference (ICSI) approach of Khalaf and Saunders (2014) enhanced to consider non-Gaussian distributions. Simulations show that the IIE leads to relatively unbiased parameter estimates, and the ICSI method leads to good coverage properties both under the null and alternative.

Commercial bank costs are analyzed in a dynamic technical efficiency model framework. A negative dynamic relationship is indicated by MLE, while IIE provides evidence that a static model is more appropriate. The confidence intervals are constructed under the ICSI method, which show that the model errors have a positive skew.

The remainder of the paper is structured as follows. Section 6.2 presents the current models available in the literature, and present the dynamic stochastic frontier model introduced in this paper. Section 6.3 presents the MLE approach and introduces the IIE and ICSI methods. Section 6.4 covers the Monte Carlo simulation study, which demonstrates the MLE-bias and IIE bias-correction, and the coverage properties of the ICSI method to construct confidence sets. Section 6.5 examines cost inefficiency of commercial banks under the assumption of dynamic technical inefficiency. Section 6.6 summarizes and provides direction to future research.

6.2 Methods

The dynamic technical efficiency model put forth in Ahn and Sickles (2000) is defined by

$$y_{i,t} = x_{i,t}\beta + (\beta_0 + \gamma t) + v_{i,t} - u_{i,t}, \quad (6.1)$$

$$u_{i,t} = (1 - \rho_i)u_{i,t-1} + (1 - \rho_i)\gamma + \eta_{i,t}, \quad (6.2)$$

where $v_{i,t}$ are i.i.d. symmetric errors, $\eta_{i,t} \geq 0$, and $0 < \rho_i \leq 1$, such that $u_{i,t} \geq 0$. To take into account the serial correlation of the errors, the model can be re-specified as:

$$y_{i,t} = (1 - \rho_i)y_{i,t-1} + (x_{i,t} - (1 - \rho_i)x_{i,t-1})\beta \quad (6.3)$$

$$+ \rho_i\beta_0 + \rho_i\gamma t + (v_{i,t} - (1 - \rho_i)v_{i,t-1}) - \rho_i\eta_{i,t}. \quad (6.4)$$

Clearly the lagged dependent series is correlated with $v_{i,t-1}$, so the authors suggest two solutions for estimation. The first assumes that $v_{i,t} = (1 - \rho_i)v_{i,t-1} + h_{i,t}$ where $h_{i,t}$ are i.i.d. symmetric errors centered at zero, and the parameters can be estimated via non-linear GLS. The second uses appropriate lags of the dependent series as instruments to estimate the parameters by non-linear GMM.

The authors test the assumption that individual efficiency dynamics are equal, $\rho_i = \rho$ for all i , which places the model closer to the standard dynamic panel frameworks of Arellano and Bond (1991), Ahn and Schmidt (1995), Arellano and Bover (1995), and Blundell and Bond (1998). The GMM approach provides consistent parameter estimates, but more recent studies have found that these estimates are susceptible to weak instruments, especially near the unit root boundary (see Phillips 2014), and a variance ratio of the errors and unobserved heterogeneity that is not unity (see Bun and Windmeijer 2010).

A second representation of dynamic technical efficiency model put forth in Desli et al (2003) is

$$y_{i,t} = x_{i,t}\beta + \alpha_{i,t} + v_{i,t} \text{ and} \quad (6.5)$$

$$\alpha_{i,t} = \alpha_i + \rho\alpha_{i,t-1} + w_{i,t}\gamma - u_{i,t}, \quad (6.6)$$

where $u_{i,t} \geq 0$, and $w_{i,t}$ are a set of technical efficiency specific covariates. The model can be represented as a panel-ARMA, specifically

$$y_{i,t} = \alpha_i + \rho y_{i,t-1} + (x_{i,t} - \rho x_{i,t-1})\beta + w_{i,t}\gamma + \epsilon_{i,t} \text{ and} \quad (6.7)$$

$$\epsilon_{i,t} = (v_{i,t} - \rho v_{i,t-1}) - u_{i,t}. \quad (6.8)$$

This representation will suffer from an incidental parameter bias due to the correlation of the unobserved heterogeneity (α_i) and the lagged dependent parameter ($y_{i,t-1}$).

The third representation of dynamic technical efficiency is that of Tsionas (2006), given as

$$y_{i,t} = x_{i,t}\beta + v_{i,t} - u_{i,t} \text{ and} \quad (6.9)$$

$$\ln(u_{i,t}) = \rho \ln(u_{i,t-1}) + w_{i,t}\gamma + \eta_{i,t}, \quad (6.10)$$

where $v_{i,t} \sim IN(0, \sigma_v^2)$ and $\eta_{i,t} \sim IN(0, \sigma_\eta^2)$. Although less clear by this representation, if any regressor in $x_{i,t}$ is a time-invariant regressor, for example x_i , will be correlated with the lagged errors ($u_{i,t-1}$).

The dynamic technical efficiency model of both Desli et al (2003) and Tsionas (2006) will lead to biased parameter estimates. Without loss of generality, this paper focuses on an autoregressive panel without regressors, specifically

$$y_{i,t} = \rho y_{i,t-1} + (1 - \rho)\alpha_i + \eta_{i,t} \text{ and} \quad (6.11)$$

$$\eta_{i,t} \sim \text{skew-}N(\mu, \sigma_\eta, \lambda), \quad (6.12)$$

where μ is the location, σ_η is the scale, and λ is the shape parameter (the normal distribution is recovered when $\lambda = 0$). By transformation, the errors can be alternatively defined by $\eta_{i,t} = v_{i,t} - u_{i,t}$, where $v_{i,t} \sim IN(0, \sigma_v^2)$ and $u_{i,t} \sim IN^+(0, \sigma_u^2)$.

This dynamic technical efficiency model has four parameters of interest, defined as $\theta = \{\rho, \mu, \sigma_\eta, \tau\}$. Dynamic panel models with normal errors lead to provably location-scale invariant estimates of ρ , refer to Khalaf and Saunders (2014). In contrast, panel models with skew-normal errors lead to location-scale invariant estimates of ρ , but these estimates are not shape parameter invariant.

6.3 Unbiased Estimates and Confidence Set

The indirect inference estimator (IIE) produces unbiased estimates of the parameter (θ). The parameter confidence set is constructed by inverting the indirect inference objective function (IIOF) by introducing an additional layer of Monte Carlo replications.

The estimates from the data are obtained by maximum likelihood from the demeaned series,¹

$$\eta_{i,t}^* = y_{i,t}^* - \rho y_{i,t-1}^*, \quad (6.13)$$

where the individual likelihood is given by

$$L(\eta_{i,t}^*; \mu, \sigma_\eta, \lambda) = \frac{2}{\sigma_\eta} \phi\left(\frac{\eta_{i,t}^* - \mu}{\sigma_\eta}\right) \Phi\left(\frac{\lambda(\eta_{i,t}^* - \mu)}{\sigma_\eta}\right), \quad (6.14)$$

and the total log-likelihood is

$$\mathcal{L}(\rho, \mu, \sigma_\eta, \lambda) = \sum_{i=1}^N \sum_{t=2}^T \ln(L((y_{i,t}^* - \rho y_{i,t-1}^*); \mu, \sigma_\eta, \lambda)). \quad (6.15)$$

The MLE parameter estimate of ρ can be shown to be equivalent to the fixed-effect estimator for dynamic panel models. For fixed-T, Nickell (1981) derived the analytical bias for dynamic panel models under the assumption of i.i.d. errors and shows that the bias is $O(T^{-1})$. The i.i.d. errors covers the skew-normal case, but the variance of these errors depend on both the scale (σ_η) and shape (λ) parameters. The structure of the fixed-effect estimator requires that only *strictly* exogenous regressors can be included, but these regressors can be correlated with the unobserved heterogeneity. As shown in Phillips and Sul (2007), the source of the Nickell bias is due to the correlation of the demeaned lagged dependent series with the demeaned errors, and the parameter estimates for the exogenous regressors are subsequently biased from the original Nickell bias.

The IIE is introduced as a correction approach. Let $\theta = (\rho, \mu, \sigma_\eta, \lambda)'$ represent the vector of parameters. The IIE method constructs a set of simulated series under

¹The demeaning subtracts the sample mean from the series, and is indicated by an asterisk.

a given null, $H_0 : \theta = \theta_s$, and each series is indexed by $h = 1, \dots, H$. The IIOF is defined as

$$Q(\theta_{s,0}) = \left(\hat{\theta}_0 - \frac{1}{H} \sum_{h=1}^H \hat{\theta}_h(\theta_s) \right)' \left(\hat{\theta}_0 - \frac{1}{H} \sum_{h=1}^H \hat{\theta}_h(\theta_s) \right), \quad (6.16)$$

where $\hat{\theta}_0$ is the estimate from the data. The minimization of $Q(\theta_{s,0})$ over a suitable parameter space, say $\theta_s \in \Theta$, leads to unbiased and consistent estimate of θ .

The confidence set for θ is constructed using the indirect confidence set inference (ICSI) approach, which generates an additional layer of Monte Carlo simulations that are used to calibrate the IIOF statistic. These Monte Carlo simulations are generated under the joint null, $H_0 : \theta = \theta_s$. The ICSI method uses a M -dimension set of IIOF, specifically

$$Q(\theta_{s,m}) = \left(\hat{\theta}_m - \frac{1}{H} \sum_{h=1}^H \hat{\theta}_h(\theta_s) \right)' \left(\hat{\theta}_m - \frac{1}{H} \sum_{h=1}^H \hat{\theta}_h(\theta_s) \right), \quad (6.17)$$

to simulate the empirical distribution of IIOF as a statistic. The rank of $Q(\theta_{s,0})$ in the set of Monte Carlo simulated IIOF is used to compute the Monte Carlo p -value at θ_s . The confidence set is then constructed by collecting all values of θ_s that are not rejected under the null,

$$CI(\theta) = \{\theta_s | p\text{-value}(Q(\theta_{s,0})) > \alpha\}. \quad (6.18)$$

6.4 Simulations

A small simulation study is used to examine the bias correction and the coverage of the confidence set. The data generating process for the simulations are described by Eqs. (6.11) and (6.12). The simulation study considers static ($\rho = 0$) and stationary ($\rho = 0.4$) panel settings, with symmetric errors ($\lambda = 0$) and positive skew errors ($\lambda = 2$), and for all simulations the scale parameter is fixed, $\sigma_\eta = 1.2$.

Table 6.1 presents the parameter bias ($\hat{\theta} - \theta$) of each estimation method. The lagged dependent parameter is downward biased for the MLE estimate in all settings and the IIE approach provides suitable bias correction.

The bias of distribution related parameters depends on whether it is Gaussian or not under the null. For normally distributed errors, MLE and IIE approaches lead to relatively unbiased shape parameter estimates. However, the scale parameter estimates are downward biased for MLE and upward biased for IIE. The IIE estimates under normal errors are provably location-scale invariant, see Khalaf and Saunders (2014). This is an example of a Davies (1977, 1987) problem, specifically when the true errors are normally distributed then under the null of normality the scale parameter is not identified.

Table 6.1 Estimation bias for a dynamic stochastic frontier model

			MLE			IIE		
ρ	σ_η	λ	ρ	σ_η	λ	ρ	σ_η	λ
0	1.2	0	-0.249	-0.105	0.011	0.013	0.175	-0.032
0	1.2	2	-0.239	-0.294	-0.721	0.014	0.000	-0.018
0.4	1.2	0	-0.376	-0.139	0.011	0.014	0.196	-0.016
0.4	1.2	2	-0.365	-0.340	-0.810	0.016	0.005	0.039

Based on 1000 replications

Table 6.2 ICSI rejection frequency for lagged dependent parameter

ρ	{ $\rho = 0.4, \lambda = 0$ }	{ $\rho = 0.4, \lambda = 2$ }
-0.8	0.705	1.000
-0.6	0.386	1.000
-0.4	0.209	0.998
-0.2	0.114	0.939
0.0	0.067	0.418
0.2	0.055	0.104
0.4	0.044	0.050
0.6	0.050	0.081
0.8	0.062	0.221

Based on 1000 replications, $\sigma_\eta = 1.2$, significance level of 5 %, and size indicated in bold

Table 6.3 ICSI rejection frequency for shape parameter

λ	{ $\rho = 0.4, \lambda = 0$ }	{ $\rho = 0.4, \lambda = 2$ }
-5	1.000	1.000
-2	0.978	1.000
-1	0.278	0.998
0	0.044	0.806
1	0.259	0.082
2	0.975	0.050
5	1.000	0.731

Based on 1000 replications, $\sigma_\eta = 1.2$, significance level of 5 %, and size indicated in bold

The IIE parameter estimates are unbiased when the true errors are non-Gaussian, and the scale parameter is identified under both the null and alternative. The MLE parameter estimates are downward biased for all parameters, for positive skew errors.

Tables 6.2 and 6.3 present the rejection frequency properties of the ICSI method for a dynamic technical efficiency model. The former table tests under the null of various values of ρ and the latter for various shape parameter values (the null used in the data generating processes are indicated in bold). The rejection frequency under the null is size corrected with rejection frequencies close to the desired 5 % level. The inversion of the IIOF statistic in the ICSI framework ensures that the Monte Carlo and II simulations are exchangeable under the null.

The rejection frequency under the alternative increases as the parameter setting under the null moves away from the true parameter value. A nonzero shape parameter improves power in ρ and the ICSI approach is robust to asymmetric confidence intervals for λ . The standard MLE-based confidence bounds impose symmetry, which may be a strong assumption in the context of asymmetric distributions.

6.5 Banks, Productivity, and Dynamics

The bank cost data analyzed in Kumbhakar and Tsionas (2005) is used in this study to illustrate the effect of introducing dynamics into a panel model with skew-Normal errors. The data was originally obtained from the Federal Reserve Bank of Chicago, specifically the commercial bank and bank holding company database. The database is balanced with 500 commercial banks identified, and for the years 1996 through to 2000.²

The model under consideration is a dynamic translog model with skew-normal errors, given by:

$$c_{i,t} = c_{i,t-1}\rho + X'_{i,t}\beta + (1 - \rho)\alpha_i + \eta_{i,t} \text{ and} \quad (6.19)$$

$$\eta_{i,t} \sim \text{skew-}N(\mu, \sigma_\eta, \lambda), \quad (6.20)$$

where $c_{i,t}$ is the log costs of the bank, $X_{i,t}$ is the vector of exogenous regressors (including interaction terms), and α_i is the unobserved bank-specific heterogeneity. The exogenous regressors include the quantity and prices of labour, capital, purchased funds, interest-bearing deposits in total transactions accounts, and interest-bearing deposits in total nontransaction accounts.

The exogenous regressors are partialled out to focus on the bias-correction of the lagged dependent parameter, and the parameters of the skew-normal distribution. A projection matrix, M_X is constructed and under the assumption of *strict* exogeneity.³ The series are demeaned (identified with an asterisk), so the matrix representation of the model is:

$$M_X c^* = M_X c_{-1}^* \rho + M_X \eta^* \text{ and} \quad (6.21)$$

$$M_X \eta^* \sim \text{skew-}N(\mu, \sigma_\eta, \lambda), \quad (6.22)$$

²Data set downloaded from the website of William Greene.

³On the definition of M_X : Gouriéroux et al (2010) recommend the usual OLS-based projection matrix; we use an alternative form proposed and justified by Saunders (2015).

Table 6.4 Dynamic translog model of commercial banks

			IIE 95 % Confidence interval	
	MLE	IIE	Lower	Upper
ρ	-0.2340	0.0123	-0.4620	0.4786
σ_η	0.2564	0.3676	<0.00001	0.9052
λ	1.5449	3.1886	0.7536	>15.0000

The demeaning of the series forces the sample and expected mean of the estimation errors to be exactly zero. The estimate of μ is obtained recursively by:

$$\hat{\mu} = -\hat{\sigma}_\eta \frac{\hat{\lambda}}{\sqrt{1 + \hat{\lambda}}} \sqrt{\frac{2}{\pi}}, \quad (6.23)$$

since it is exactly defined from the other parameters then it is excluded from the objective functions of the estimators.

Let $\theta = \{\rho, \sigma_\eta, \lambda\}$ represent the model parameters that are estimated by MLE and IIE. The MLE estimates are obtained by maximizing the total likelihood given by:

$$\mathcal{L}(\rho, \sigma_\eta, \lambda) = \sum_{i=1}^N \sum_{t=2}^T \ln(L((M_X c^* - M_X c_{-1}^* \rho)_{i,t}; \mu(\sigma_\eta, \lambda), \sigma_\eta, \lambda)). \quad (6.24)$$

The MLE estimates are used as the estimates from the data (θ_0) for the IIE and ICSI methods outlined above (Table 6.4).

The maximum likelihood estimates indicate that a negative relationship between the previous period's realization of costs and the current period and that the errors are positively skewed. The IIE parameter estimates indicate that there is no dynamic relationship between periods for the bank's costs. The more prominent positive skew for the IIE approach matches the upward bias correction observed in Table 6.1.

The confidence intervals, constructed by the ICSI approach, fails to reject $\rho = 0$, which coincides with a static translog model. This is important, since the MLE approach indicates that the dynamic parameter is significant due to the downward bias of MLE. The shape parameter, λ , does not cover zero, which indicates that the normal distribution is not covered by the confidence set.

6.6 Final Remarks

We show that the IIE and ICSI approaches provide promising bias-correction and coverage for dynamic stochastic frontier models. Our analysis is seen as a first step to correct for known estimation bias in similar models. The proposed IIE approach uses MLE as the auxiliary estimator, similar to the fixed-effects estimator, while alternative estimators may improve efficiency of the ICSI method.

The time-varying technical (in)efficiency of the estimation is recoverable, in the manner described in Desli et al (2003), but the proposed method allows for more elaborate error structure even in the presence of biased-MLE estimation. The unobserved heterogeneity in the proposed framework can be recovered recursively based on the unbiased parameter estimates. The IIE framework allows for non-Gaussian and Gaussian distribution assumptions on the unobserved heterogeneity.

References

- Ahn SC, Schmidt P (1995) Efficient estimation of models for dynamic panel data. *J Econ* 68(1):5–27
- Ahn SC, Sickles RC (2000) Estimation of long-run inefficiency levels: a dynamic frontier approach. *Econ Rev* 19(4):461–492
- Arellano M, Bond S (1991) Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *Rev Econ Stud* 58(2):277–297
- Arellano M, Bover O (1995) Another look at the instrumental variable estimation of error-components models. *J Econ* 68(1):29–51
- Blundell R, Bond S (1998) Initial conditions and moment restrictions in dynamic panel data models. *J Econ* 87(1):115–143
- Bun MJ, Windmeijer F (2010) The weak instrument problem of the system gmm estimator in dynamic panel data models. *Econ J* 13(1):95–126
- Davies RB (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64(2):247–254
- Davies RB (1987) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 74(1):33–43
- Desli E, Ray SC, Kumbhakar SC (2003) A dynamic stochastic frontier production model with time-varying efficiency. *Appl Econ Lett* 10(10):623–626
- Gouriéroux C, Phillips PC, Yu J (2010) Indirect inference for dynamic panel models. *J Econ* 157(1):68–77
- Khalaf L, Saunders CJ (2015) Confidence intervals in autoregressive panels, invariance, and indirect inference. Manuscript
- Kumbhakar SC, Tsionas EG (2005) Measuring technical and allocative inefficiency in the translog cost system: a bayesian approach. *J Econ* 126(2):355–384
- Nickell S (1981) Biases in dynamic models with fixed effects. *Econ J Econ Soc* 49:1417–1426
- Phillips PC (2014) Dynamic panel gmm with roots near unity. Technical Report, Working Paper, Yale University
- Phillips PC, Sul D (2007) Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence. *J Econ* 137(1):162–188
- Saunders CJ (2015) On autoregressive panels with covariates: invariance, indirect inference, and confidence sets. Manuscript
- Tsionas EG (2006) Inference in dynamic stochastic frontier models. *J Appl Econ* 21(5):669–676

Chapter 7

Analysing Labour Productivity in Ecuador

German Cubas, Anson T. Y. Ho, Kim P. Huynh, and David T. Jacho-Chávez

Abstract We studied labour productivity growth in Ecuador from 1998 to 2006 by using firm-level data from the annual survey of manufacturing and mining. This period is characterised by the economic crisis in 1999 and important economic reforms. During the crisis, there was a two percent annual decrease in productivity in 1998–2000, but the recovery was strong with a five percent annual productivity growth in 2002–2004. Our productivity decomposition indicated that the main source of productivity growth came from firms with increasing productivity gaining market shares. Within-firm productivity decline was substantial during the crisis, but its growth was secondary in the post crisis recovery. Firm entry and exit only had minor impacts on labour productivity. Our distributional analysis further showed that labour productivity distribution increased in 2000–2002 and had remained at higher level for the rest of the sample period.

Keywords Labour productivity • Decomposition • Distributional analysis

JEL codes: F31, D24, L11

G. Cubas (✉)

Department of Economics, University of Houston, 204 McElhinney Hall, Houston,
TX 77204-5019, USA

e-mail: gcubasnorando@uh.edu

A.T.Y. Ho

Department of Economics, Kansas State University, 327 Waters Hall, Manhattan,
KS 66506-4000, USA

e-mail: atyho@ksu.edu

K.P. Huynh

Bank of Canada, 234 Wellington Ave., Ottawa, ON, Canada K1A 0G9
e-mail: kim@huynh.tv

D.T. Jacho-Chávez

Department of Economics, Emory University, Rich Building 306, 1602 Fishburne Dr.,
Atlanta, GA 30322-2240, USA

e-mail: djachocha@emory.edu

7.1 Introduction

Emerging economies are characterised by the severity of their recessions. Ecuador is an interesting case study for small open developing economy, because its economic crisis in 1999 was triggered by negative shocks (for details, see Beckerman 2002; Jácome 2004) and consequently important economic reforms were carried out. It serves as a natural experiment for us to investigate the impacts of deep recession on firm survival and productivity. In this paper we focused on labour productivity because it is often used and does not require estimating a production function using less reliable firm-level capital data.

Using the annual survey of manufacturing and mining from 1998 to 2006, we found that about 25 % of firms exited in 5 years. Firms entered during the crisis had higher survival rates than those entered in subsequent years. During the crisis period in 1998–2000, labour productivity decreased by 1.9 % per year. The recovery was slow at the beginning with productivity growth of 1.1 % per year in 2000–2002. It increased dramatically to 5.1 % in 2002–2004. After 2004, productivity growth slowed down to 2 %. On average, labour productivity increased by 1.6 % per year from 1998 to 2006.

To understand the source of labour productivity growth, we conducted a productivity decomposition proposed by Foster et al. (2001) (hereafter FHK). We found that, first, labour productivity growth within continuing firms (*within* effect) contributed –7.7 % annual productivity growth in 1998–2000 and remained at about 2 % afterwards. Second, firms that became more productive also gained market shares (*cross* effect). This effect was the major source of productivity growth, which overall contributed 2.1 % annual productivity growth. Third, firms with initially above-average labour productivity had lost market share (*between* effect). On average, it contributed –1.2 % annual productivity growth. This apparent negative *between* effect is noted by Petrin et al. (2013) for the case of Chile. In our study, the combination of *cross* and *between* effect indicated that changes in market shares among continuing firms made positive contributions to labour productivity growth. These effects were strongest during the crisis and in the immediate recovery period, which implies there was substantial reallocation of market share due to changes in firm's productivity. Last, *net entry* only had minor contribution to productivity growth during the sample period.

In addition to analysing the aggregate productivity growth, we further investigated the evolution of productivity distribution using the functional principal components analysis suggested by Kneip and Utikal (2001). We found large increase in labour productivity distribution first 2 years after the crisis (2000–2002), and our statistical test rejects the hypothesis that labour productivity distribution has remained the same throughout the time period.

This paper is organised in the following fashion: Sect. 7.2 describes the data used and offers firm survival statistics. Section 7.3 performs the labour productivity decomposition. Section 7.4 discusses the distributional analysis on the labour productivity. Section 7.5 concludes.

7.2 Data

Our analyses are based on firm-level data from 1998 to 2006 reported in the annual survey of manufacturing and mining (Encuesta Anual de Manufactura y Minería). It is a survey of firms with at least 10 employees published by Instituto Nacional de Estadística y Censos (INEC). The data was cleaned to maintain longitudinal consistency. Output is defined as value-added sales and the change in inventory, deflated to US\$2000 using Ecuadorian sectoral producer price index. Labour is defined as the total number of paid workers. Firm-level labour productivity is calculated as firm output divided by the number of paid workers. Firms with top and bottom 1 % of labour productivity were excluded.

Table 7.1 reports the conditional survival rates. Firms were grouped into cohorts based on their observed entry year. Since we did not observe the entry year for firms that already existed in 1998, they were grouped in the same bin as pre-1998 cohort. On average, about 25 % of firms exited in 5 years. First-year exit rate differed due to the timing of entry. Firms entered during the crisis had higher survival rates than those entered in subsequent years. The higher survival rates can be rationalised as only strong firms entered the market; Huynh et al. (2010) find a similar result for Canada. However, we did not find any noticeable change in survival rates in the recovery years after the crisis.

7.3 Labour Productivity Decomposition

Foster et al. (2001) suggest the following decomposition (hereafter FHK decomposition) to document how restructuring within an entry cohort contributes to overall changes in labour productivity. The output-weighted labour productivity (\hat{P}_{ct}) for cohort c in year t is

Table 7.1 Survival rates

Cohort	1998	1999	2000	2001	2002	2003	2004	2005	2006
Pre-1998	1.00	0.90	0.84	0.79	0.74	0.69	0.64	0.61	0.58
1999	.	1.00	0.94	0.88	0.82	0.76	0.71	0.68	0.65
2000	.	.	1.00	0.93	0.87	0.80	0.75	0.71	0.68
2001	.	.	.	1.00	0.93	0.85	0.80	0.76	0.72
2002	1.00	0.92	0.84	0.80	0.77
2003	1.00	0.92	0.87	0.83
2004	1.00	0.94	0.88
2005	1.00	0.94
2006	1.00

Note: Survival rate refers to the proportion of firms that continued their operations from the previous year

$$\hat{P}_{ct} = \sum_{i \in S_{ct}} s_{it} p_{it}$$

where p_{it} is labour productivity, s_{it} is the output share of firm i , and c is the particular cohort. The FHK decomposition accounts for overall changes ($\Delta \hat{P}_{ct}$) from birth until time t for a given cohort. This change can be decomposed into the following components:

$$\begin{aligned} \hat{P}_{ct} - \hat{P}_{c1} = & \underbrace{\sum_{i \in S_{ct}} s_{i1} (p_{it} - p_{i1})}_{\text{Within}} + \underbrace{\sum_{i \in S_{ct}} (p_{i1} - \hat{P}_{c1}) (s_{it} - s_{i1})}_{\text{Between}} + \underbrace{\sum_{i \in S_{ct}} (p_{it} - p_{i1}) (s_{it} - s_{i1})}_{\text{Cross}} \\ & + \underbrace{\sum_{i \in N_{ct}} s_{it} (p_{it} - \hat{P}_{c1})}_{\text{Entry}} - \underbrace{\sum_{i \in D_{ct}} s_{i1} (p_{i1} - \hat{P}_{c1})}_{\text{Exit}} \end{aligned}$$

where S_{ct} denotes continuing cohort firms at time t , N_{ct} denotes firms entering at or prior to time t , and D_{ct} denotes firms within a cohort that exit at or prior to time t . The *within* term in this decomposition shows the contribution to overall change from within by surviving firms. The *between* component captures the contribution related to changing shares, weighted by the deviations of firm's initial productivity from the cohort's initial weighted average productivity. The *cross* term represents the covariance between changes of productivity and that of firm's output share. The last two terms provide the contributions of cohort entries and exits to overall changes, respectively.

Results for the FHK decomposition are reported in Table 7.2 and graphically presented in Fig. 7.1. Changes in labour productivity are expressed in annual values. Focusing on the overall changes from 1998 to 2006, labour productivity had increased by 1.6 % per year. The *within* term only contributed 0.3 % to labour productivity growth. The *between* term was -1.2% , showing that continuing firms with initial above-average labour productivity ($p_{i1} - \hat{P}_{c1} > 0$) lost market share, and vice versa. The *cross* term was positive and the largest among all components,

Table 7.2 FHK decomposition of labour productivity

Time	Total	Within	Between	Cross	Net Entry	Entry	Exit
Period	(1+2+3+4)	(1)	(2)	(3)	(4=5-6)	(5)	(6)
1998–2006	0.016	0.003	-0.012	0.021	0.003	0.000	-0.003
1998–2000	-0.019	-0.077	-0.037	0.093	0.003	0.014	0.011
2000–2002	0.011	0.002	-0.038	0.067	-0.019	-0.006	0.013
2002–2004	0.051	0.019	-0.006	0.028	0.010	-0.001	-0.011
2004–2006	0.020	0.021	-0.023	0.030	-0.008	-0.008	0.001

Note: The whole sample period is divided into 2-year sub periods (cohorts). Changes in labour productivity are expressed in *annual* values

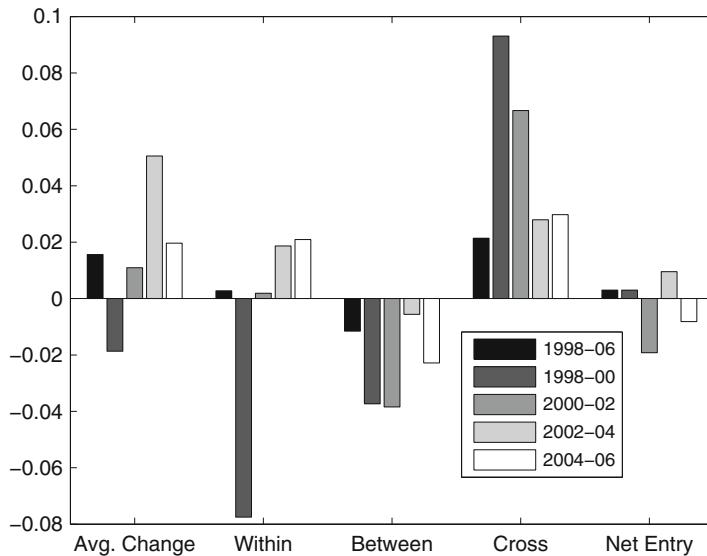


Fig. 7.1 Labour productivity FHK decomposition

contributing 2.1 % to labour productivity growth. It indicates that firms which became more productive also increased their market shares. The *between* and *cross* components imply that productivity dispersion across firms decreased over time. Firms with initially below-average productivity caught up and gained market shares from their competitors. This change in productivity distribution is evident in the bottom panel of Fig. 7.2 and formally analysed in the next section. Overall, changes in market shares among continuing firms made positive contributions to productivity growth. To investigate the importance of reallocation on productivity growth, Petrin and Levinsohn (2012) develop a productivity decomposition explicitly account for firm's optimal use of inputs, see Ho et al. (2015) for a complete analysis of Ecuador. The effect of *net entry* was close to zero since both *entry* and *exit* terms were small.

Constructing these labour productivity changes in four 2-year windows allows us to provide an understanding of which time period contributed to most of the variation. Ecuador experienced positive labour productivity growth in all sub-periods, except 1998–2000 in which the crisis hit. The negative growth during the crisis was mainly attributed to the large negative *within* firm effect of -7.7 % per year. Note that the *cross* term in 1998–2000 was strongest in all sub-periods, showing that firms with increasing productivity during the crisis expanded rapidly in market share. The recovery of labour productivity was slow at the beginning, with only 1.1 annual percent growth in 2000–2002. The *within* term was close to zero, while the weakened *cross* component was partially offset by the *between* component. *Net entry* also reduced productivity by 1.9 %, due to exits of productivity firms. Productivity growth in 2002–2004 was very strong with 5.1 % per year. It was driven by both positive *within* and *cross* terms, and the *between* term was closer to zero.

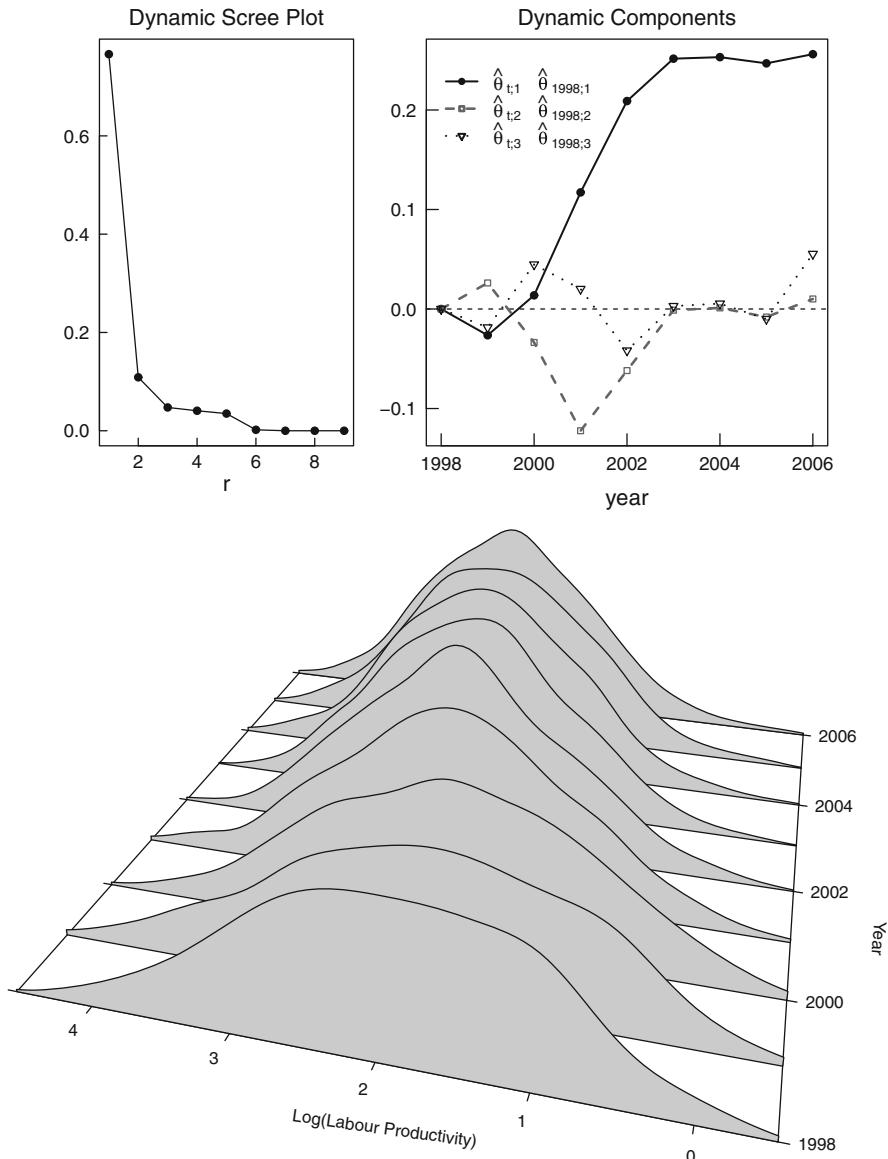


Fig. 7.2 Estimated dynamic strength components & dynamic scree plot

Note: The *top left panel* displays the Dynamic Scree plots, with each plot showing $\hat{\lambda}_r / \sum_{r=1}^T \hat{\lambda}_r$. The *top right panel* displays the estimated dynamic strength of components, with each plot showing $\hat{\theta}_{t,r} - \hat{\theta}_{t_0,r}$ for $r = 1, 2, 3$. The *bottom panel* shows the estimated distributions of labour productivity from 1998 to 2006

Overall the *between* term was negative and the *cross* term was positive in all sub-periods, with stronger effects observed between 1998 and 2002. At 2-year windows, *Net entry* still had relatively minor contributions to labour productivity growth when compared to other components.

7.4 Distributional Analysis

In this section, we studied the temporal evolution of labour productivity distributions using the functional principal component analysis (hereafter FPCA) suggested by Kneip and Utikal (2001). FPCA is a nonparametric method that allows us to identify the distributional dynamics with minimal assumptions by using their expansions into the first L functional principal components, g_1, g_2, \dots, g_L , and represent each labour productivity distribution f_t in terms of the model

$$f_t = f_\mu + \sum_{j=1}^L \theta_{t;j} g_j, \quad (7.1)$$

where $f_\mu = \sum_{t=1998}^{2006} f_t / 10$ is the common mean distribution and $L \leq 10$ corresponds to the number of non-zero eigenvalues ($\lambda_1, \lambda_2, \dots$) of the empirical covariance operator. Notice that (7.1) implies that each f_t can be obtained by adding to f_μ a transformation of compromising time-invariant common components g_1, g_2, \dots, g_L , with time-varying strengths encapsulated in the coefficients $\theta_{t;j}$. Since f_t represent densities obtained for each time period $t = 1998, \dots, 2006$, then the time evolution of their respective coefficients $\theta_{t;1}, \theta_{t;2}, \dots, \theta_{t;L}$ provides information about the evolution of the main differences and similarities between the underlying distributions. For other applications of FPCA in firm dynamics we refer the reader to Huynh and Jacho-Chávez (2010), Huynh et al. (2011), and Huynh et al. (2015).

We constructed estimators, $\hat{\lambda}_1, \hat{\lambda}_2, \dots$, and $\hat{\theta}_{t;1}, \hat{\theta}_{t;2}, \dots$, as described in Kneip and Utikal (2001). The results are summarised in Fig. 7.2. The top left panel displays the plot of $\hat{\lambda}_r / \sum_{r=1}^{10} \hat{\lambda}_r$ for labour productivity. These types of plots are known as Scree plots, but due to the time-series dimension we call them Dynamic Scree plots. The obvious observation is that the first eigenvalue dominates the scree plot, and three eigenvalues basically account for almost all the variation in the data.

Now we proceed to analyse the dynamics of $\{\hat{\theta}_{t;1}\}_{t=1998}^{2006}$, $\{\hat{\theta}_{t;2}\}_{t=1998}^{2006}$, and $\{\hat{\theta}_{t;3}\}_{t=1998}^{2006}$ through time. Top right panel of Fig. 7.2 illustrates the evolution of the deviations of the $\hat{\theta}$ series relative to their initial values in 1998. Results of the estimated dynamic strength components for labour productivity are clear. The first component showed a dramatic increase in 2001–2002 and stabilised at similar higher levels since then. The second and third component are transitory. The large increase in the first component coincides with the increase in labour productivity distribution, as illustrated in the bottom panel of Fig. 7.2.

To understand whether these FPCA components imply any statistically significant changes in the productivity distribution, we performed a test that all densities $\{f_t\}_{t=1998}^{2006}$ are equal, i.e. $H_0 : L = 0$ as suggested in Kneip and Utikal (2001, Sect. 2.3, pp. 522–523). They proposed using $\hat{\rho} = \sum_{r=1}^L \hat{\lambda}_r$ as a plausible test statistic, as well as a bootstrap procedure for approximating its distribution under the null hypothesis. Performing this test for the distribution of labour productivity yielded a bootstrapped p -value of zero. This result indicates that the hypothesis that labour productivity distribution has stayed the same through time can be rejected.

7.5 Conclusion

Using Ecuadorian firm level data from 1998 to 2006, we documented firm survival statistics and conducted an empirical analysis on labour productivity to understand the impacts of a large economic crisis in a small open developing economy. We found that the economic crisis and consequent reforms had important impacts on labour productivity. Labour productivity decreased by 2 % per year during the crisis, and it increased by 5 % per year in the later recovery period (2002–2004). The major contributor to productivity growth was the positive covariance between firms' labour productivity and their market shares. Our result from functional principal component analysis also shows that labour productivity distribution had changed through time. The findings in this paper motivate further research on changes in market friction during severe economic recession.

Acknowledgements We acknowledge the use of the np package by Hayfield and Racine (2008) and the use of the Quarry High Performance Cluster at Indiana University, where all the computations were performed. We thank Paul Carrillo, John Earle, Dan Kovenock, Jamil Mahuad, Mitsukuni Nishida, and participants of the 2010 Canadian Economic Association meetings and 2014 North American Productivity Workshop. The staff at the Ecuadorian National Statistics Office (INEC) provided great assistance with the data and answered many of our queries, in particular Galo Arias, Telmo Molina, Verónica Velázquez, Margarita Viera, and Byron Villacís. Ho thanks the U.S. Census Bureau for conference financial support. Jacho-Chávez thanks the Center for Latin American and Caribbean Studies (CLACS) at Indiana University for financial support. The views stated herein are those of the authors and not necessarily of the Bank of Canada. All errors are our own.

References

- Beckerman P (2002) Longer-term origins of ecuador's predollarization crisis. In: Beckerman P, Solimano A (eds) Crisis and dollarization in Ecuador. The World Bank, Washington
- Foster L, Haltiwanger JC, Krizan CJ (2001) Aggregate productivity growth: lessons from microeconomic evidence. In: Dean E, Harper M, Hulten C (eds) New developments in productivity analysis. University of Chicago Press, Chicago, pp 303–372

- Hayfield T, Racine JS (2008) Nonparametric econometrics: the np package. *J Stat Softw* 27(5):1–32. <http://www.jstatsoft.org/v27/i05/>
- Ho ATY, Huynh KP, Jacho-Chávez DT (2015) Productivity, reallocation, and distortions: evidence from ecuadorian firm-level data. Mime
- Huynh KP, Jacho-Chávez DT (2010) Firm size distributions through the lens of functional principal components analysis. *J Appl Econ* 25(7):1211–1214. <http://ideas.repec.org/a/jae/japmet/v25y2010i7p1211-1214.html>
- Huynh KP, Jacho-Chávez DT, Petrunia RJ, Voia M (2011) Functional principal component analysis of density families with categorical and continuous data on Canadian entrant manufacturing firms. *J Am Stat Assoc* 106(495):858–878. <http://ideas.repec.org/a/bes/jnlasa/v106i495y2011p858-878.html>
- Huynh KP, Jacho-Chávez DT, Petrunia RJ, Voia M (2015) A nonparametric analysis of firm size, leverage and labour productivity distribution dynamics. *Empir Econ* 48(1):337–360
- Huynh KP, Petrunia RJ, Voia M (2010) The impact of initial financial state on firm duration across entry cohorts. *J Ind Econ* 58(3):661–689. <http://ideas.repec.org/a/bla/jindex/v58y2010i3p661-689.html>
- Jácome LI (2004) The late 1990s financial crisis in ecuador: institutional weakness, fiscal rigidities, and financial dollarization at work. IMF working paper WP/04/12
- Kneip A, Utikal KJ (2001) Inference for density families using functional principal component analysis. *J Am Stat Assoc* 96(454):519–542
- Petrin A, Levinsohn J (2012) Measuring aggregate productivity growth using plant-level data. *Rand J Econ* 43(4):705–725
- Petrin A, Nishida M, Polanec S (2013) Explaining reallocation's apparent negative contribution to growth. NBER working paper 19012, National Bureau of Economic Research

Chapter 8

Hierarchical Performance and Unobservable Heterogeneity in Health: A Dual-Level Efficiency Approach Applied to NHS Pathology in England

A.S.J. Smith, J. Buckell, P. Wheat and R. Longo

Abstract Understanding the source of inefficiency within health system organisational structures is a key aspect of performance measurement and management; and is of increasing importance to policy makers. This study uses a unique panel dataset to study the efficiency performance of pathology services in the National Health Service (NHS) in England for the first time. We apply a dual-level stochastic frontier (DLSF) model (*J Prod Anal* 37(1):1–27, 2012) to isolate the source of inefficiency at two vertically distinct organisational levels: an upper level of Strategic Health Authorities (SHAs); and a lower level of laboratories grouped within SHAs. We develop the DLSF framework—in line with recent developments in the wider panel data literature—to control for the influence unobserved heterogeneity, which is a key issue for healthcare performance analysis. We find statistically significant variation in inefficiency performance at both organisational levels in pathology services. We use these measures to compute overall inefficiency for NHS pathology services, and corresponding savings estimates. Finally we comment on the wider modelling implications of our research with respect to the separation of inefficiency and unobserved heterogeneity when applied to multi-level data structures.

Keywords Stochastic frontier analysis • Dual-level efficiency • Unobserved heterogeneity • Costs • Regional variation • Pathology

A.S.J. Smith

Institute for Transport Studies, University of Leeds, Leeds, LS2 9LJ, UK

Leeds University Business School, University of Leeds, Leeds, LS2 9LJ, UK

J. Buckell (✉) • R. Longo

Academic Unit of Health Economics, University of Leeds, 2.07 Charles Thackrah Building, Clarendon Road, Leeds, LS2 9LJ, UK

e-mail: ts10jasb@leeds.ac.uk

P. Wheat

Institute for Transport Studies, University of Leeds, Leeds, LS2 9LJ, UK

8.1 Introduction

The National Health Service (NHS) in England is under substantial and growing financial pressure to reduce costs, through improving efficiency, whilst maintaining the quality of services (NHS England 2013; Appleby et al. 2014). Comparative efficiency analysis is therefore important as it pinpoints where performance is sub-optimal and identifies examples of best practice. Through setting targets based on the results of comparative efficiency assessment policy makers can incentivise organisations to improve in order to meet their objectives through adopting best practice.

An important aspect of measuring performance is being able to locate the source of inefficiency. This allows initiatives to adopt best practice to be targeted effectively. In health markets, organisations—particularly the NHS—are typified by hierarchical managerial structures where inefficiency may arise at different points within the (vertical) hierarchy as well as horizontally between organisational units at the same organisational level.

Recent health literature has begun to recognise that organisational structure should be incorporated into performance analysis (Adams et al. 2003; Olsen and Street 2008; Sørensen et al. 2009; Castelli et al. 2013; Zhang et al. 2013). However, the previous health efficiency literature focuses its attention on horizontal comparisons, albeit at different levels of aggregation depending on the study (see Murillo-Zamorano and Petraglia 2011; D'Amico and Fernandez 2012; Felder and Tauchmann 2013).

In this paper, we carry out a multi-level efficiency analysis that seeks to identify where inefficiency resides within a vertical organisational hierarchy in NHS pathology services. Pathology services are conducted in laboratories providing diagnostic medicine to primary care (local GPs) and secondary care (hospitals) within the NHS. We focus on pathology for two reasons, as noted by Hollingsworth (2008): first, focussing on a speciality within health services (as opposed to broader entities such as whole hospitals or regions as the unit(s) of analysis) is more likely to be useful to policy makers and thus likely to encourage the use of efficiency predictions; and second, efficiency studies in health should target specific policy objectives: this study feeds into the NHS's 'a call to action' for efficiency improvements (NHS England 2013).

Pathology services are organised hierarchically, where groups of laboratories are under the direction of Strategic Health Authorities¹ (SHAs hereafter). SHAs dictate central policy, corporate culture and have some degree of control over pathology services (e.g. the configuration of services) (Department of Health 2006); leaving some managerial autonomy at laboratory level. Thus, there is a component of overall inefficiency attributable to each SHA (which is persistent across laboratories

¹The NHS has recently undergone a substantial reorganisation under which the SHAs have been abolished. However, they were in place during the period under study.

within the SHA). Lower down in the organisation, inefficiency is likely to vary according to the relative ability of laboratory-level management. From a policy and management perspective, it is important to understand both sources of inefficiency so that appropriate incentives can be offered to drive improvements in efficiency. By combining the inefficiency estimates from the two hierarchical levels (persistent and lab-varying), an overall measure of inefficiency for the higher level (SHA) can be computed. Measuring multi-level performance may be of greater practical use than single level measures, which should encourage their uptake amongst policy makers, which has hitherto been limited in health markets (Hollingsworth 2012).

To obtain inefficiency measures at these different organisational levels, and an overall inefficiency measure, we adopt the dual level stochastic frontier model (DLSF; see Smith and Wheat (2012)), which has been applied in other sectors to measure multi-level firm inefficiency. The advantage of this model is firstly that it enables inefficiency at different organisational levels to be identified. Smith and Wheat (2012) use the terminology sub-company, or internal inefficiency, which in our case corresponds to inefficiency at the laboratory level; and persistent or external inefficiency, which in our case refers to persistent inefficiency at the SHA level. A further key finding of their paper is that, when the organisational structure is not accounted for, inefficiency predictions can exhibit a downward bias. Thus there is motivation in adopting a DLSF model both to yield insight into the level of inefficiency variation at different levels and to eliminate bias in the overall prediction.

Another form of bias, which is particularly problematic in health, results from the failure to appropriately model unobservable heterogeneity (Greene 2004; Smith et al. 2012). In the case of pathology services, there are significant differences in laboratories' production processes. These may include factors typically studied by economists such as outputs and input prices; but also specificities such as patient mix or service quality, *inter alia* (Department of Health 2006; Buckell et al. 2013; NHS England 2014; Buckell et al. 2015). Some of these features are difficult or even impossible to measure directly, and so accounting for unobservable heterogeneity is of paramount importance.

Smith and Wheat (2012) recognise that persistent inefficiency at the higher level of aggregation (which corresponds, in our case, to SHA level inefficiency) could also reflect time and laboratory invariant unobserved heterogeneity. However, they leave that issue for future research. At the same there has been considerable interest in the wider panel data stochastic frontier literature (Farsi et al. 2005a; Kumbhakar et al. 2014) on how to separate inefficiency from unobserved, time invariant heterogeneity. We therefore augment the Smith and Wheat (2012) DLSF approach to reflect developments in the wider panel data efficiency literature with regard to the vexing problem of disentangling inefficiency from unobserved heterogeneity. We compare our findings to a model without any attempt to separate inefficiency and unobservable heterogeneity, to demonstrate the importance of accounting for the latter case of multi-level data structures.

Our paper therefore contributes to the literature in two ways. It is the first application of the Smith and Wheat (2012) DLSF inefficiency model in a health context. It is also the first time the approaches set out in Farsi et al. (2005a) and Kumbhakar et al. (2014) have been applied to a multi-level data structure. We thus apply and develop state-of-the-art models to draw policy conclusions on pathology services within the NHS in England, and also offer insights on the relative merits of different approaches to separating inefficiency and unobserved heterogeneity when applied to multi-level data structures.

The remainder of this paper is as follows. In Sect. 8.2, our models and estimation strategy are discussed. Section 8.3 details the data. Section 8.4 presents our results and discusses. Section 8.5 concludes.

8.2 Methodology

We begin our methodological discussion with the general form of the dual-level stochastic frontier proposed by Smith and Wheat (2012). We next discuss the issue of unobservable heterogeneity and its relevance to efficiency estimation in our application. We then outline two further models, each of which adds a component to the model to distil out the unobservable heterogeneity from the efficiency prediction, though with differing assumptions. We finally consider a fully generalised model comprising the features of the preceding models. In total, four models are considered. Estimation, econometric specification and statistical testing are described.

Our starting point is the Dual-Level Stochastic Frontier (DLSF) model proposed by Smith and Wheat (2012). This model is derived from panel data stochastic frontier models, with the exception that the structure of the panel is amended from firm and time to firm and sub-company, where the sub-company units are repeat observations of their respective firms. In this way, the structure of the organisation is embodied in the model. This allows the decomposition of inefficiency at the two organisational levels in the hierarchy. In this application, SHAs are equivalent to firms, and laboratories are equivalent to the sub-company units.

Smith and Wheat (2012) outline the advantages of this model and its application to multi-level data structures. First, multi-level data structures increase the number of observations for analysis, which can be a major benefit for economic regulators who often have to work with small cross-sections and limited time periods. Second, it permits a clearer understanding of where inefficiency resides in the vertical hierarchy, allowing regulators to target the elimination of persistent differences between SHAs (external inefficiency) and differences in performance of laboratories within the same SHA (internal inefficiency). Finally, it is beneficial to conduct performance analysis at the level of disaggregation that relates to how SHAs/laboratories actually organise themselves, in particular allowing the true scale properties of the cost function to be established.

The imposed form of inefficiency is well suited to the multi-level model. Smith and Wheat (2012) note that, in traditional panels, having an overall inefficiency comprising a component of SHA inefficiency that is time-invariant and a component that varies randomly over time may not accurately capture the natural temporal evolution of inefficiency. In contrast, imposing a SHA-invariant component and a laboratory-varying component to the structure of inefficiency befits the aim of vertically decomposing inefficiency.

The DLSF model takes the general form,

$$C_{i,s} = \delta_i + X'_{i,s}\beta + \varepsilon_{i,s} \quad (8.1)$$

$$\varepsilon_{i,s} = \tau_{i,s} + v_{i,s} \quad (8.2)$$

$$\tau_{i,s} \sim N^+(0, \sigma_\tau^2) \quad (8.3)$$

$$v_{i,s} \sim N(0, \sigma_v^2) \quad (8.4)$$

where $C_{i,s}$ is the cost of laboratory s in SHA i . $X_{i,s}$ is a vector of outputs, input prices and environmental variables; β is a vector of parameters to be estimated. δ_i is the SHA-specific effect. $\tau_{i,s}$ is laboratory-specific inefficiency and $v_{i,s}$ is random statistical noise. The notation in (8.1) highlights the tiered structure of the data only; in the empirical work presented below, there is also a time dimension to the data.

Estimation proceeds via estimation of a SHA-stratified random effects model (REM) by Generalised Least Squares (GLS) (as in (8.1)), yielding estimates of β ($\hat{\beta}$), predicted values of SHA effects (to which we turn our attention in the following sections), and residuals, $\hat{\varepsilon}_{i,s}$.

The prediction of laboratory-specific inefficiency is conducted in a second stage. It is common for all four models. We take the model residuals from the first stage (which have had the SHA effect removed), stratify by laboratory and apply the Jondrow et al. (1982) procedure to retrieve laboratory-specific predictions of inefficiency,

$$\hat{\tau}_{i,s} = E[\tau_{i,s} | \tau_{i,s} + v_{i,s}] \quad (8.5)$$

We assume time-invariance for the predicted efficiency at laboratory level, given that our panel is both short in its time dimension and unbalanced. The competing models are then distinguished according to the treatment of the SHA-specific effect, δ_i .

8.2.1 *The Dual Level Stochastic Frontier (Model 1)*

The DLSF treats the SHA-specific effect as inefficiency, which in the case of GLS estimation yields,

$$\delta_i = \alpha_0 + \mu_i \quad (8.6)$$

$$\mu_i \sim N\left(0, \sigma_\mu^2\right) \quad (8.7)$$

where the prediction of SHA inefficiency is a Schmidt and Sickles (1984)-type correction, $\widehat{\mu}_i = \widehat{\mu}_i - \min(\widehat{\mu}_i)$.

8.2.2 *Accounting for Unobservable Heterogeneity*

A simplifying assumption of the DLSF proposed by Smith and Wheat (2012) was that the SHA effect is interpreted as the SHA inefficiency. This is consistent with the received literature such as Kumbhakar and Heshmati (1995). However, Smith and Wheat (2012) acknowledged that this interpretation may not be appropriate in all cases. In particular, any heterogeneity that is not captured by the regressors is incorporated into this effect, which biases inefficiency estimates (Kumbhakar and Lovell 2000). Ultimately, the SHA effect is a mixture of unobservable effects, one of which being SHA invariant inefficiency.

In the case of pathology production, there are features of laboratories' production environments for which no data are available, e.g. the service quality (which is known to vary between laboratories and SHAs), implying the DLSF may be an inappropriate specification. As such, the DLSF model is extended to examine two approaches to incorporate the influence of unobservable heterogeneity, namely the use of the Mundlak (1978) transformation and the residual decomposition approach of Kumbhakar et al. (2014). In addition, we estimate a model which incorporates both of these approaches. We utilise statistical testing to determine an appropriate approach. Results are compared between models to demonstrate differences.

8.2.3 *The Mundlak-Transformed DLSF (Model 2)*

One way to introduce a control for unobservable heterogeneity into the DLSF model follows Farsi et al. (2005a), which was first extended to the DLSF by Wheat (2014). The approach makes use of Mundlak's (1978) recognition of the link between

random and fixed effects in panel data models. This approach is operationalised via a direct insertion of group means of the regressors into the random effects model.² In this way, this model nests model 1.

This model assumes that inefficiency is uncorrelated with the regressors whilst unobserved heterogeneity is assumed to be correlated with the regressors. Correlation between the SHA effects and the regressors is modelled explicitly by using the variable group means. Under the assumption that this correlation represents unobservable heterogeneity, it is removed from the SHA effects. Then the SHA effects that remain are treated as before and efficiency predictions are derived.

$$\delta_i = \alpha_0 + \bar{X}'_i \rho + \mu_i \quad (8.8)$$

$$\mu_i \sim N\left(0, \sigma_\mu^2\right) \quad (8.9)$$

Here, $\bar{X}'_i \rho$ captures unobservable heterogeneity that is correlated with the regressors. SHA inefficiency predictions, $\hat{\mu}_i$, are: $\hat{\mu}_i = \hat{\mu}_i - \min(\hat{\mu}_i)$.

Model 2 has a number of appealing features. First, the separation of inefficiency from unobservable heterogeneity (that is correlated with the regressors) is achieved. Second, consistent, unbiased within estimators for the frontier parameters are recovered through application of GLS to this model.³ Third, it is possible to examine the relationship between the unobservable heterogeneity and the variables via the group mean coefficients ($\hat{\rho}_{GLS}$) (Farsi et al. 2005a, b). Fourth, this model does not require any additional stages; the model is estimated exactly as the DLSF with the addition of the group mean variables. Fifth, the restriction (no correlation between the regressors and unobserved heterogeneity) can be readily tested using a Wald test on the joint hypothesis: $\rho = 0 \forall \bar{X}_i$ (which is referred to as the Wu test (Greene, 2008)).

There are some drawbacks to using this method. First, the model relies on the assumption that the unobservable heterogeneity is correlated with the regressors while inefficiency is assumed to be completely uncorrelated with regressors. Thus any unobservable heterogeneity that is uncorrelated with regressors is interpreted as inefficiency and, conversely, any inefficiency correlated with regressors (but firm invariant) is interpreted as unobserved heterogeneity. Finally, relative to the simpler DLSF, the Mundlak transformation proliferates parameters, which will reduce the precision of parameter estimates.

²There is an alternative approach using a fixed effects model and an auxiliary regression on the SHA effects (Farsi et al. 2005a). In linear models, this method returns identical parameter estimates, but underestimates standard errors in the auxiliary stage, so the random effects approach is preferred (see Baltagi 2006, p. 1192, for the variance of the group means in the REM).

³We note that within estimators are in some cases imprecise, which to some extent diminishes their appeal.

8.2.4 The Four-Component DLSF (Model 3)

A second approach to amend the DLSF to account for unobservable heterogeneity is to follow the approach of Kumbhakar et al. (2014) based on their four-component model. The application to our hierarchical data is similar to model 1, except for an additional stage to separate the firm inefficiency from the unobservable heterogeneity (the latter now assumed uncorrelated with the regressors). In this way, model 3 nests model 1.

In this additional stage, the SHA effects are decomposed by imposing distributional assumptions and applying a stochastic frontier to them. Thus, unobservable heterogeneity is assumed to embody the features of statistical noise in traditional stochastic frontiers (SFs) ((8.10)–(8.13) below). Unobservable heterogeneity is assumed to be uncorrelated with the regressors. This is in direct contrast to the Mundlak approach (Wheat 2014). Here, SHA inefficiency is computed using the Jondrow et al. (1982) method, rather than the Schmidt and Sickles (1984) approach used in models 1 and 2.

$$\delta_i = \alpha_0 + \mu_i \quad (8.10)$$

$$\mu_i = \alpha_i + w_i \quad (8.11)$$

$$\alpha_i \sim N^+ (0, \sigma_\alpha^2) \quad (8.12)$$

$$w_i \sim N (0, \sigma_w^2) \quad (8.13)$$

where w_i represents unobserved heterogeneity that is uncorrelated with the regressors and inefficiency is calculated as: $\hat{\alpha}_i = E[\alpha_i|\alpha_i + w_i]$.

The benefits of this model are that, firstly, it is possible to control for unobservable heterogeneity. Second, it is possible to test the decomposition of the inefficiency and the unobserved heterogeneity by applying routine tests in the SF literature. Third, although full distributional assumptions are made to predict inefficiency, the parameter estimates of the frontier are estimated using much weaker (and thus robust) assumptions in the first stage, which is a noteworthy advantage over a single stage alternative (Smith and Wheat (2012)).

There are disadvantages to implementing this model. First, relative to the simple DLSF, there are additional assumptions on the error components necessary to enable separation of inefficiency from unobserved heterogeneity, and these are arbitrary. Second, the SF procedure to obtain SHA persistent inefficiency predictions is conducted on the number of SHAs, which may be small in empirical applications (in our case 10); and, in turn, may yield imprecise parameter estimates, particularly with respect to the variances of the SHA invariant error components. In traditional panels it is also the case that this part of the procedure faces limitations if the

cross-section is small. In addition, the multi-stage approach yields standard errors of second stage parameter estimates smaller than their true magnitude owing to the use of first stage residuals in the second stage (Kumbhakar et al. (2014) note that this issue is routinely disregarded). However the fundamental limitation of this approach is the assumption that unobserved heterogeneity is uncorrelated with the regressors, which in turn requires reliance on distributional assumptions to separate inefficiency from the unobserved heterogeneity.

8.2.5 *The Mundlak-Transformed Four Component DLSF (Model 4)*

Our final model is a DLSF that is augmented for unobservable heterogeneity by combining the three approaches above. In this model, inefficiency is purged of both types of unobserved heterogeneity, that is, unobserved heterogeneity that is correlated with the regressors, and that which is not. Thus the appeal of this specification is that the somewhat restrictive assumptions about the correlation between unobservable heterogeneity and the regressors in the two prior approaches can be (a) relaxed and (b) tested. We therefore specify the following,

$$\delta_i = \alpha_0 + \bar{X}'_i \rho + \mu_i \quad (8.14)$$

$$\mu_i = \alpha_i + w_i \quad (8.15)$$

$$\alpha_i \sim N^+(0, \sigma_\alpha^2) \quad (8.16)$$

$$w_i \sim N(0, \sigma_w^2) \quad (8.17)$$

where $\bar{X}'_i \rho$ captures unobservable heterogeneity that is correlated with the regressors and w_i represents unobserved heterogeneity that is uncorrelated with the regressors. Inefficiency is calculated as: $\hat{\alpha}_i = E[\alpha_i | \alpha_i + w_i]$.

Model 4 nests its component models—it is possible to test down to arrive at a preferred model. In particular, it is possible to test each of the components individually, and examine the presence and/or form of unobservable heterogeneity, and to remove it from the estimates of inefficiency.

Overall, four models are estimated and tested: the dual level stochastic frontier (DLSF) of Smith and Wheat (2012) (Model 1); the DLSF with the Mundlak adjustment applied (Model 2); the four-component DLSF model based on Kumbhakar et al. (2014) (Model 3); and the Kumbhakar-DLSF model with the Mundlak adjustment applied (Model 4). Table 8.1 below shows the econometric specifications of these models.

Table 8.1 Econometric specifications of models 1–4

Model	Stage 1: RE GLS	SHA inefficiency	Laboratory (sub-company) inefficiency
(1)	$C_{i,s} = \alpha_0 + X'_{i,s}\beta + \varepsilon_{i,s}$	$\widehat{\mu}_i = \widehat{\mu}_i - \min(\widehat{\mu}_i)$	$\widehat{\tau}_{i,s} = E[\tau_{i,s} \tau_{i,s} + v_{i,s}]$
(2)	$C_{i,s} = \alpha_0 + \bar{X}'_i\rho + X'_{i,s}\beta + \varepsilon_{i,s}$	$\widehat{\mu}_i = \widehat{\mu}_i - \min(\widehat{\mu}_i)$	$\widehat{\tau}_{i,s} = E[\tau_{i,s} \tau_{i,s} + v_{i,s}]$
(3)	$C_{i,s} = \alpha_0 + X'_{i,s}\beta + \varepsilon_{i,s}$	$\widehat{\alpha}_i = E[\alpha_i \alpha_i + \omega_i]$	$\widehat{\tau}_{i,s} = E[\tau_{i,s} \tau_{i,s} + v_{i,s}]$
(4)	$C_{i,s} = \alpha_0 + \bar{X}'_i\rho + X'_{i,s}\beta + \varepsilon_{i,s}$	$\widehat{\alpha}_i = E[\alpha_i \alpha_i + \omega_i]$	$\widehat{\tau}_{i,s} = E[\tau_{i,s} \tau_{i,s} + v_{i,s}]$

From Table 8.1, we note that stage 1 is identical for models 1 and 3; and for models 2 and 4. In model 1 and model 2, the predicted SHA inefficiencies are derived from $\widehat{\mu}_i$. In models 3 and 4, the SHA effects are decomposed to yield inefficiency predictions according to the distributional assumptions specified.

Laboratory inefficiency predictions for models 1 and 3 are identical as a corollary of the common first stage. Similarly, models 2 and 4 have identical predicted laboratory inefficiencies.

We now turn the choice between models 1–4. We are able to use statistical tests to guide model selection. Table 8.2 summarises our model testing. We first test the SHA effects using a Moulton-Randolph test (a Standardised Lagrange Multiplier test (SLM)), which is better suited to unbalanced panels (as in our panel) than the standard LM test (Moulton and Randolph 1989). We then move to testing the decomposition of inefficiency and unobservable heterogeneity. The unobservable heterogeneity that is correlated with the regressors is tested using a Wald test on the group mean variables jointly. This test is applied to models 2 and 4. To test unobservable heterogeneity that is uncorrelated with the regressors, we use a LR test on the SHA SF. These tests apply to models 3 and 4. Finally, the test of the presence of inefficiency at the laboratory level is tested using a LR test on the laboratory level SF.

8.2.6 Overall Efficiency

Finally, having retrieved the two efficiency predictions at the separate hierarchical levels, it is necessary to compute an overall efficiency for the SHA—our persistent, top-level inefficiency measure—which is the sum of its SHA-specific inefficiency and the (cost) weighted average of its constituent laboratories’ inefficiencies. We use this measure to compute our overall savings estimates. Taking model 1 as an example,

$$\bar{u}_i = \widehat{\mu}_i + \frac{\sum_{\forall s} C_{i,s} \cdot \widehat{\tau}_{i,s}}{\sum_{\forall s} C_{i,s}} \quad (8.18)$$

Table 8.2 Statistical tests on models 1–4

	Model 1	Model 2	Model 3	Model 4
<i>Test of firm effects</i>				
Firm effects (vs. pooled model)	Moulton-Randolph H_0 : no firm effects	Moulton-Randolph H_0 : no firm effects	Moulton-Randolph H_0 : no firm effects	Moulton-Randolph H_0 : no firm effects
<i>Decomposition</i>				
Inefficiency and UOH correlated with regressors		Wald test on \bar{X}_i H_0 : $\rho = 0 \forall \bar{X}_i$		Wald test on \bar{X}_i H_0 : $\rho = 0 \forall \bar{X}_i$
Inefficiency and UOH uncorrelated with regressors			LR of firm effect SF H_0 : no inefficiency	LR of firm effect SF H_0 : no inefficiency
<i>Test of laboratory inefficiency</i>				
Test of sub-company inefficiency	LR on sub-company SF H_0 : no inefficiency	LR on sub-company SF H_0 : no inefficiency	LR on sub-company SF H_0 : no inefficiency	LR on sub-company SF H_0 : no inefficiency

UOH Unobserved heterogeneity, LR likelihood ratio, SF stochastic frontier

8.3 Data

Annual pathology benchmarking data is used to compile an unbalanced panel of 57 English NHS pathology laboratories amongst ten Strategic Authorities during the 5 year period from 2006/7 to 2010/11. The sample represents approximately one third of the 163 NHS pathology laboratories in England.

Our dependent variable is the laboratory's total operating costs (net of capital charges).

Output is measured by the number of requests for tests. We could, of course, use the number of tests actually carried out as our output measure. However, laboratories are known to conduct varying numbers of tests per request, which may distort the measure of output if it is based on tests. We further capture this variation by including a variable defined as the ratio of tests to requests (variable name Tests:Requests), in addition to our output measure. Input prices for labour are based on data from the UK labour force survey. Labour force survey data is chosen over other sources (NHS staff census data, for example) to ensure the exogeneity of the data.⁴ In the absence of other input prices data, this variable is considered a proxy for labour and materials.

Variables capturing exogenous characteristics include: a binary variable for the foundation status of the host trust,⁵ meaning that it has financial autonomy (variable name Foundation). It is expected that foundation status trusts will have lower operating costs than their non-foundation counterparts owing to a more commercial outlook towards service provision (Healthcare commission 2007). We also include a binary variable (variable name Metropolitan) denoting within an urban area or city; the null case is rural. This is to capture the differences in service provision between rural and urban patient populations and their differing pathology demands, e.g. a broader range of diseases in larger cities (Department of Health 2006).

Descriptive statistics are presented in Table 8.3. Costs and wage data are in real terms (2007 prices), adjusted using the consumer prices index (CPI). The ratio of tests to requests is calculated from the data, as are variable group means for the

Table 8.3 Descriptive statistics

Variable	Mean	S.D.	Min	Max
Operating costs (adjusted)	3,617,320	2,058,358	963,875	11,741,895
Number of tests	5,037,362	2,990,846	1,380,384	30,199,502
Number of requests	714,125	465,535	191,078	4,423,531
Input prices (labour) (adjusted)	24,551	4160	15,834	49,955

⁴Mutter et al. (2013) demonstrate using healthcare data that endogeneity can bias efficiency scores.

⁵The term 'trust' in the NHS refers to a single hospital or a small group of hospitals in close proximity (e.g. in an urban area) which operate as a single entity.

Mundlak transformation. For estimation, natural logarithms of variables are taken. We use a Cobb-Douglas functional form.⁶ LIMDEP software is used for estimation (Greene 2012a, b).

8.4 Results and Discussion

In this section our results are summarised and discussed. We begin with our parameter estimates from the first stage of models 1–4. Next, we discuss model selection and select our preferred model. We then move to the efficiency predictions and our savings estimates. Finally, we comment on the health policy and wider modelling implications of our empirical results.

8.4.1 Parameter Estimates

Models 1–4 use a random effects model as the first stage in estimation. Models 2 and 4 extend the model with the Mundlak group mean variables. Therefore, two model outputs are reported: one with a Mundlak adjustment (models 2 and 4), and one without a Mundlak adjustment (models 1 and 3). Table 8.4 reports the model outputs.

Table 8.4 shows the parameter estimates from both of the first stage models. The $\hat{\beta}$ are similar between models and similar to findings in other studies of pathology services (Buckell et al. 2013, 2015).⁷ The within estimators do not appear to exhibit imprecision (which was a concern of adopting this approach, see Sect. 8.2).

In both models the output coefficients are positive and significant, suggesting, at the sample mean, increasing returns to scale (RTS) properties in pathology production (since $RTS = 1/\hat{\beta}_{output} = 1/0.897 = 1.115$). This corresponds to results and/or predictions from other pathology studies (Department of Health 2006, 2008; Healthcare Commission 2007; Holland and Trigg 2011).

We find that laboratories facing higher input prices have higher costs; that laboratories with higher tests-to-requests ratios have higher operating costs; and that laboratories in urban settings have higher operating costs (coefficient on the Metropolitan variable), which is in agreement with other pathology studies (Department of Health 2006). The within estimator suggests that the foundation variable is not significant, whilst this variable is found to be statistically significant at the 5 % level in the REM without Mundlak. The study of the Healthcare Commission (2007) suggested that the foundation of the host trust may lead to lower operating costs, although no empirical results were presented.

⁶We tested a Translog specification, however, the coefficients on some key variables were not significant. Therefore, we prefer a Cobb-Douglas specification which gives a credible set of parameter estimates, and a more credible model from which our efficiency predictions are derived.

⁷We note that similar data is used for these studies so this result is not surprising.

Table 8.4 Model outputs for Mundlak adjusted and non-Mundlak adjusted random effects models

	REM with Mundlak			REM without Mundlak		
	Model 2 Model 4			Model 1 Model 3		
	Beta	s.e.	Sig	Beta	s.e.	Sig
Constant	1.285	5.497		-5.833	1.712	***
Output (requests)	0.897	0.043	***	0.897	0.043	***
Input prices	0.892	0.161	***	0.774	0.153	***
Tests:Requests	0.549	0.066	***	0.547	0.069	***
Metropolitan	0.196	0.046	***	0.198	0.047	***
Foundation	-0.065	0.041		-0.081	0.041	**
Time	-0.021	0.012	*	-0.019	0.129	
REQBAR	-0.334	0.194	*			
INPBAR	-0.287	0.451				
TESBAR	-1.070	0.552	*			
METBAR	0.097	0.203				
FOUBAR	0.222	0.169				
TIMBAR	0.234	0.130	*			

*, **, *** Denote statistical significance at the 10 %, 5 % and 1 % level, respectively. s.e.—standard errors. The Mundlak group mean variables are denoted “XXXBAR” and correspond to their respective variables above

The coefficient on the time variable, representing technical change (frontier shift), is significant only in the REM with Mundlak (the within estimator). The coefficient suggests that pathology costs are, on average across the market, decreasing annually by around 2 % owing to technical change. This finding is in keeping with the empirical findings of Holland and Trigg (2011). Moreover, this result is intuitively sound, as a heavily mechanised industry such as pathology is likely to be characterised by technological change over time, leading to cost reductions, even in the short run (as in this data).

We now discuss the Mundlak group mean coefficients. There appears to be divided opinion in the literature as to their interpretation individually, although most authors do not comment on them in isolation. Of those that do, Farsi et al. (2005a, b) take the view that the group means indicate correlation between the variable and unobservable heterogeneity. Conversely, Filippini and Hunt (2012) state that the interpretation of these variables is not straightforward, and do not assign any interpretation to these coefficients. In our application, we are interested in the decomposition of efficiency and unobservable heterogeneity, thus the interpretation of these variables is of no specific interest to us.

Overall, the Mundlak-transformed model is considered a better reflection of the economic reality than its non-transformed counterpart on a priori grounds as it permits inefficiency estimates to be purged of unobserved heterogeneity that is correlated with the regressors. We discuss model selection based on appropriate statistical testing below.

Table 8.5 Test statistics, models 1–4

	Model 1	Model 2	Model 3	Model 4
<i>Firm effects (vs. pooled model)</i>				
Moulton-Randolph	2.696***	4.168***	2.969***	4.168***
<i>Decomposition of inefficiency and unobserved heterogeneity</i>				
Wald test of 6 linear restrictions, $\rho = 0 \forall \bar{X}_i$		12.89**		12.89**
LR of firm effect SF (vs. OLS)		0		1.147
<i>Test of sub-company inefficiency</i>				
LR of laboratory SF (vs. OLS)	64.689***	58.170***	64.689***	58.170***

*, **, *** Denote statistical significance at the 10 %, 5 % and 1 % level, respectively

8.4.2 Model Selection

We now move to our discussion on model selection. To begin, we consider the testing procedure outlined in Table 8.2. We discuss the results from these tests, which are reported in Table 8.5. We also draw on the model efficiency predictions, which are presented in Table 8.6.

The first issue is whether the multi-level structure is appropriate. From the significant Moulton-Randolph statistic, the panel specification of the first stage formulation is preferred to the pooled model, supporting the presence of SHA effects. Of course, as noted above, we then need to consider the interpretation and decomposition of these SHA effects.

For all models, the LR statistic on the laboratory level SF is significant, supporting the presence of inefficiency at laboratory level.

We now turn to the unobservable heterogeneity test statistics. The Wald test of six linear restrictions—the Wu test—indicates that the variable group means are jointly statistically significant additions to the model.⁸ There is thus evidence to support the correlation between the SHA effects and the regressors, which we interpret as unobservable heterogeneity. On this basis, we prefer model 2 to model 1 and model 4 to model 3.

As expected, model 1 appears to confound unobservable heterogeneity with inefficiency. This issue is well known in the health context (Greene 2004; Farsi et al. 2005a). When the Mundlak adjustment is applied, the average predicted efficiency increases significantly from 0.625 to 0.715 (Table 8.6). This finding, combined with the results of the Wu test, suggests that there is a substantial amount of unobservable heterogeneity that is correlated with the regressors.

⁸We have also used the more familiar Hausman test. In this case, however, the test statistic could not be computed because the variance-covariance matrix is not positive definite. We thus revert to the Wu test (Greene 2012b) and note that, in any case, reliance on the Hausman statistic alone is discouraged (Baltagi 2008).

Table 8.6 Efficiency predictions at SHA level, laboratory level and overall efficiency with overall ranks, models 1–4

	Model 1			Model 2			Model 3			Model 4						
	SHA	SHA	Lab	Overall	Rank	SHA	Lab	Overall	Rank	SHA	Lab	Overall	Rank			
A	0.902	0.726	0.655	2	0.907	0.734	0.666	9	1.000	0.726	0.726	9	0.983	0.734	0.721	5
B	1.000	0.827	0.827	1	1.000	0.814	0.814	1	1.000	0.827	0.827	1	0.987	0.814	0.803	1
C	0.800	0.785	0.628	3	0.888	0.791	0.703	5	1.000	0.785	0.785	5	0.905	0.791	0.716	6
D	0.797	0.772	0.615	6	0.974	0.782	0.762	2	1.000	0.772	0.772	6	0.978	0.782	0.765	2
E	0.669	0.805	0.538	10	0.813	0.808	0.657	10	1.000	0.805	0.805	2	0.882	0.808	0.712	7
F	0.755	0.767	0.579	8	0.952	0.772	0.735	4	1.000	0.767	0.767	7	0.964	0.772	0.744	4
G	0.727	0.786	0.571	9	0.920	0.799	0.735	3	1.000	0.786	0.786	4	0.935	0.799	0.748	3
H	0.841	0.736	0.619	5	0.936	0.739	0.691	7	1.000	0.736	0.736	8	0.951	0.739	0.702	10
I	0.792	0.786	0.622	4	0.871	0.793	0.691	8	1.000	0.786	0.786	3	0.889	0.793	0.705	8
J	0.825	0.719	0.593	7	0.952	0.730	0.695	6	1.000	0.719	0.719	10	0.964	0.730	0.704	9
Mean	0.811	0.771	0.625		0.921	0.776	0.715		1.000	0.771	0.771		0.944	0.776	0.732	
s.d.	0.092	0.035	0.078		0.054	0.031	0.047		0.000	0.035	0.035		0.039	0.031	0.033	

Model 3 is unable to detect any inefficiency at the SHA level (Table 8.6)—the SHA effects exhibited wrong skew. As noted, the Wu test result suggests that there is a high amount of unobservable heterogeneity that is correlated with the regressors, which model 3 does not allow for. Therefore, the finding of zero inefficiency is likely more a matter of model misspecification than of economic reality. This finding suggests that controlling for unobservable heterogeneity that is correlated with the regressors is vital: had we estimated only models 1 and 3, we might have concluded that there is no inefficiency at the SHA level and that SHA effects were driven by heterogeneity. We therefore prefer model 2 to models 1 and 3.

The final model selection decision is then a choice between model 2 and model 4. This choice hinges on the result of the attempt to decompose the SHA effect into inefficiency and unobserved heterogeneity that is correlated with the regressors (stage 2 in model 4). Although inefficiency was detected at the SHA level in model 4 (which was not the case in model 3), the result was not statistically significant (Table 8.5). The conclusion, at face value then, is that once purged of unobserved heterogeneity (correlated and uncorrelated with the regressors) there is no statistically significant SHA-level inefficiency.

However, we note that stage 2 of the multi-stage approach is based on only ten observations (as we have only 10 SHAs). As a result, the failure to find inefficiency in this model is unsurprising (this is likely to be an issue for this model on any dataset, like ours, where the number of firm observations is low).

We further note a striking concordance between the predicted SHA efficiencies of models 2 and 4 with respect to rank (Kendall's tau = 0.600**, see Table 8.6 for ranks), absolute correlation (=0.92⁹) and mean predicted efficiency (model 2 = 0.921; model 4 = 0.944, see Table 8.6). So, although the inefficiency effects are not statistically significant when making the final decomposition of inefficiency and unobserved heterogeneity (uncorrelated with the regressors), the inefficiency predictions and ranks are scarcely affected. It appears that much of the unobservable heterogeneity is correlated with the regressors and it is then difficult to disentangle the remaining effect.

Overall, we conclude that there is some remaining inefficiency at the SHA level and in the discussion that follows, we use model 4 as our preferred model. This is on the grounds that it takes account of unobserved heterogeneity that is uncorrelated with the regressors, noting that results are very similar if we were to revert to model 2.

⁹Farsi et al. (2005a, b) suggest, as a rule of thumb, any score greater than 0.9 can be considered as similar; our result is well in excess of this.

Table 8.7 Rank correlation (Kendall's tau) between overall inefficiency predictions, models 1–4

	Model 1	Model 2	Model 3	Model 4
Model 1				
Model 2	0.022			
Model 3	-0.022	0.156		
Model 4	0.156	0.600**	0.289	

*, **, *** Denote statistical significance at the 10 %, 5 % and 1 % level, respectively

8.4.3 SHA, Laboratory Level and Overall Efficiency Predictions

Table 8.6 shows the efficiency predictions from the four models. For each model there are four columns corresponding to the SHA-specific efficiency, the laboratory-specific efficiency, the overall efficiency and the rank of the SHA in terms of its overall efficiency. For the first three columns, the means of the predicted efficiencies and corresponding standards deviations are provided.

Table 8.7 shows the rank correlations between the predicted overall efficiencies for models 1–4.¹⁰ As can be seen, there is very little concordance between almost all of the models' predicted ranks. This is not entirely surprising given that model 1 makes significantly different assumptions to the remaining models and that model 3 failed to recognise any inefficiency at the SHA level. The exception to the trend is that the predicted ranks of model 2 and model 4 are statistically significantly correlated.

Model 1 exhibits the lowest predicted efficiency with a mean overall efficiency of 0.625. This is as expected given that, by construction, this model makes no allowance for the effect of unobservable heterogeneity on efficiency prediction. Thus, the unobservable heterogeneity is encompassed in the inefficiency component of the model. This issue is well known in the health context (Greene 2004; Farsi et al. 2005a).

In model 2, it is assumed that unobservable heterogeneity is correlated with the regressors. As such, we are able to use the procedure outlined in Sect. 8.2 to remove it from the SHA effects. Here, the mean overall efficiency increases significantly to 0.715.

In model 3, the unobservable heterogeneity is assumed to be uncorrelated with the regressors and assumed to embody a set of assumptions (Sect. 8.2). In this application, the SHA effects that had a SF applied to them (stage 2 of the multi-stage approach) exhibited wrong skew. Thus, no inefficiency was detected at the SHA level; that is, the firm effect is entirely composed of unobservable heterogeneity. In this sense, model 3 predicts the highest SHA efficiency. As noted, we believe this model to be misspecified.

¹⁰We use Kendall's tau to measure rank correlation, which is well suited to small samples (Kendall and Gibbons 1990).

Model 4 combines both the assumptions and procedures of the preceding three models: unobservable heterogeneity is assumed to be, in part, correlated with regressors and, in part, uncorrelated with the regressors.

As can be seen, as expected, the predicted mean overall efficiency in model 4, 0.732, is higher than that of model 2, 0.715. The difference is slight in contrast to the predictions of model 1 versus the predictions of model 2, suggesting that there is less unobservable heterogeneity that is uncorrelated with the regressors than that which is correlated with the regressors. This indicates that the Mundlak adjustment appears to capture almost the full extent of the unobservable heterogeneity. However, there was a small difference between the predicted efficiency ranks and SHA efficiencies, suggesting that the additional control is worth retaining.

There is a more fundamental point when comparing model 3 with models 2 and 4, which is that there is potential for model misspecification, which may have serious implications for findings. In our case, this could lead to what we believe to be an incorrect conclusion about the performance of the SHAs: zero inefficiency. This underlines the importance of accounting for both forms of unobservable heterogeneity discussed here.

As discussed in Sect. 8.2, the predicted laboratory efficiencies are identical in pairs: the laboratory efficiency predictions of models 1 and 3 are one pair; and of models 2 and 4 are the other pair. That is, there are two ‘sets’ of laboratory efficiency predictions. These two sets of efficiency predictions are very similar with regard to their averages, 0.771 and 0.776 (Table 8.6), their absolute correlation ($=0.98$) and their rank correlation (Kendall’s tau $= 0.956^{***}$). This suggests that efficiency predictions at the laboratory level are robust to the specification of unobservable heterogeneity (or indeed whether it is assumed away, as in model 1). This result likely arises from the similarity between the estimated model parameters in the first stage(s).

We note in passing that there may be a residual amount of unobservable heterogeneity between laboratories within SHAs; we did not investigate this issue and are not aware of models that would permit this. We therefore leave this for future research.

8.4.4 Implications for Health Policy

To begin, the overall inefficiency predicted by our model for pathology services in the NHS is around 27 % (see Table 8.6). Therefore, through appropriate target setting, it should be possible to make substantial efficiency gains in services as a whole (that is, even the best performing SHAs can improve). By overall region, the most efficient SHA is B¹¹ with an overall inefficiency of around 20 % (see (8.18) for derivation); and the least efficient region in SHA H with an overall inefficiency

¹¹Due to data confidentiality we are unable to reveal the identity of SHAs.

of 30 %. It should be noted that even the most efficient SHAs have room to improve because of variations in the laboratory performance within them (discussed below). The efficiency gap between the best and worst performing regions is around 10 %. SHAs I and J are also close to the SHA H level of inefficiency. Thus, pathology policy makers should look to these SHAs for maximum gains.

To calculate potential monetary savings, we take the efficiency prediction of each laboratory in its final year, apply its cost weight and compute the potential saving per laboratory. When this is aggregated across all of the laboratories, we find £54 m of potential annual savings in the sample. If this is applied to all NHS pathology services, this would suggest potential savings of around £675 m per annum.¹² This is significantly more than found in other empirical studies (£250–500 m in Department of Health 2008; £390 m in Buckell et al. 2015).

Next, our model enables policy makers to look within SHAs to locate the source of overall inefficiency. As envisaged at the outset, we find inefficiency at both levels, but laboratory inefficiency dominates. The mean inefficiency at the SHA level is relatively low at 6 %, where the least well performing SHA has 12 % inefficiency. In contrast, the mean inefficiency at the laboratory level is much greater at 22 %, and the least well performing group of laboratories appears to be 27 % inefficient. Thus targets and policy mechanisms would appear to be better aimed at reducing or exploring differences in performance between laboratories within SHAs, rather than looking at persistent efficiency differences between different SHAs.

A further advantage of this model is that it allows policy makers to observe inefficiency differences between individual laboratories; variation that is concealed when considering average laboratory inefficiency for each of the SHAs (which can be seen from Table 8.6 do not vary enormously). In Fig. 8.1, we see that two laboratories have inefficiency that >40 %: laboratories 12 and 38. Laboratory 38 in particular should be singled out by policy makers to improve its performance given an inefficiency of 56 %. We note that these predictions do not encompass the effects of the SHAs, which have been removed. Of course, as noted earlier, further examination of those laboratories would be needed as it may be that part of the efficiency gap is explained by other factors not taken account of in our model.

For several reasons, the use of efficiency studies by health policy makers, despite their prevalence, has been limited (Hollingsworth 2008, 2012; cf.). We have addressed three of these issues in this study. First, as is clear from our paper, this modelling framework gives a complete top-to-bottom view of pathology services. In doing so, we are able to indicate the precise location of the inefficiency in these services, which is not possible with single level approaches, making our results of greater use in a practical sense. Second, we have purposefully focussed on a speciality of health services (as opposed to more aggregated entities such as whole

¹²In our sample, we have only one third of English laboratories, none of the laboratories in Wales, Scotland or Northern Ireland and one of five pathology disciplines. We thus follow other pathology studies and apply our overall savings to total pathology expenditure to arrive at our estimate.

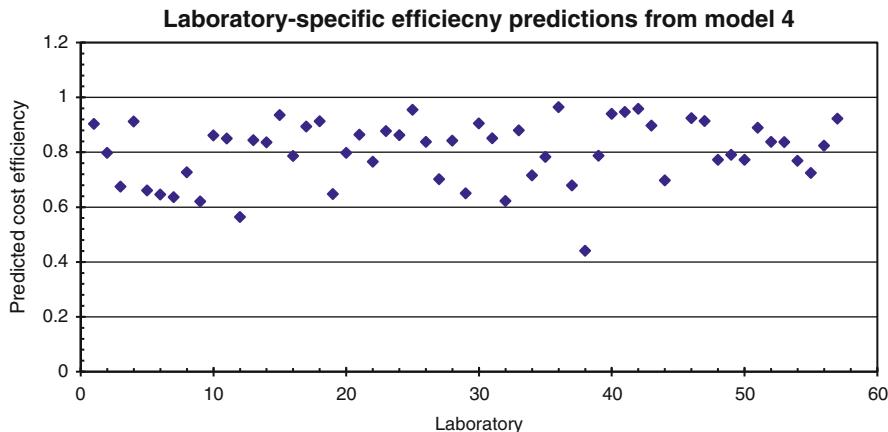


Fig. 8.1 Laboratory efficiency predictions from model 4. NB—to preserve the anonymity of the SHAs and laboratories, we do not assign the laboratories to their SHAs in this graph

hospitals or health regions)—pathology—again to make our findings of use to policy makers. Third, we have targeted a specific policy: the NHS’s “A Call To Action” to make efficiency gains (NHS England 2013).

8.4.5 Implications for Modelling Multi-Level Data Structures

We now turn to the wider modelling implications of our work. We have already noted the advantages of adopting the multi-level model as it is possible for policy makers to observe inefficiency at different levels (Smith and Wheat 2012). The alternatives, namely pooling laboratory level data, or modelling at the SHA level of aggregation do not (by construction) allow this decomposition (Smith and Wheat 2012). In preliminary analysis we estimated both of these alternatives and found that overall inefficiency was underestimated, which is in keeping with Smith and Wheat (2012). This is likely contributing to the differences in efficiency savings between our findings and other pathology studies (Sect. 8.4).

However, the contribution of this paper—apart from being the first health application of the Smith and Wheat (2012) DLSF—is to augment that model to control for unobserved heterogeneity in a multi-level context. We have shown the importance of accounting for unobserved heterogeneity in our study. This has clear implications for policy. Of course, we found that inefficiency is overestimated when unobservable heterogeneity is disregarded (which is well known in the health literature). We also found that unobservable heterogeneity arises in various forms; specifically, we find that it is important to take account of unobserved heterogeneity that is correlated with the regressors as well as that which is not. Indeed, we find

that models that do not take account of the former, such as the recently developed approach by Kumbhakar et al. (2014) (model 3 in our paper), may lead to unrealistic predictions and erroneous conclusions.

Further, in the context of multi-level structures, we noted that it may be hard to distinguish inefficiency from unobserved heterogeneity that is uncorrelated with regressors. This is because this part of the decomposition is based on the number of observations at the SHA level, which in our case is only 10. Thus there may be limits to the degree to which unobserved heterogeneity can be separated from inefficiency in data structures of this nature. As a caveat to this statement, a finding of no inefficiency when applying the Kumbhakar et al. (2014) model could be a reflection of underlying economic reality, and not necessarily because of misspecification or lack of data points (though we believe the latter to be the case in our example). Of course, the same problem, namely lack of observations to decompose inefficiency and unobserved heterogeneity that is uncorrelated with the regressors, also arises in traditional panels with a small cross-section.

8.5 Conclusions

This paper is the first application of the Smith and Wheat DLSF (2012) in a health context and the first time vertically distinct measures of inefficiency have been simultaneously estimated in health markets. It is also the first time the approaches set out in Farsi et al. (2005a) and Kumbhakar et al. (2014), to control for unobserved heterogeneity, have been applied to a multi-level data structure.

Our results suggest overall inefficiency in pathology services in England of around 27 %. This would correspond to annual savings of approximately £675 m if applied to all NHS pathology services. This estimate exceeds previous studies' savings estimates, thus suggesting the scope for further improvements than have previously been envisaged (which is a conclusion in keeping with that of other application of this model; see Smith and Wheat (2012)).

The source of the inefficiency is visible in our study, which was not the case in previous studies. The results show that the dominant source of inefficiency is variation at the laboratory level inefficiency within SHAs, though the SHA-level persistent inefficiency effects are also important. This illumination of the location of inefficiency should provide a useful guide for policy makers. Our results further show that some individual laboratories have particularly high inefficiency, which is worthy of further investigation.

With respect to the method, we find that it is important to consider both sources of unobservable heterogeneity (correlated and uncorrelated with the regressors). In our case, unobserved heterogeneity that is correlated with the regressors dominates. We note that the Kumbhkar et al. (2012) model (model 3 in our paper) did not detect SHA-level inefficiency (wrong skew), which we attribute to model misspecification given that it neglects an important source of unobserved heterogeneity. Model 4, which takes account of both sources of unobserved heterogeneity, struggled to

disentangle inefficiency from unobserved heterogeneity that is uncorrelated with the regressors. We attribute this problem to the fact that this stage of the decomposition relied on only ten observations (as we have 10 SHAs). This could be a limitation to the degree to which unobserved heterogeneity can be separated from inefficiency in data structures of this nature, where there may be a small number of observations for the top-level of the hierarchy. Of course, the same problem would occur in traditional panel models with a small cross-section. We do note, however, that failure to separate inefficiency from unobserved heterogeneity that is uncorrelated with the regressors could simply reflect economic reality rather than caused by model misspecification and/or lack of data points (though we believe the latter to be true in our example).

Whilst the different approaches produced different results for SHA-level inefficiency, the inefficiency predictions at the laboratory level were largely the same across all models. It appears, then, that the inefficiency estimates at this lower level are robust to the treatment of unobserved heterogeneity. This is likely due to the estimated parameters being similar between models. However, we consider that further research might incorporate how unobserved heterogeneity at the lower level might be incorporated into the modelling framework.

Acknowledgements The authors thank an anonymous reviewer for some useful comments. The authors also thank participants at two conferences for helpful discussion, namely the 8th North American Productivity Workshop (Ottawa, June 2014) and the European Health Policy Group meeting (Pisa, April 2014). This work was funded under an ‘innovations in quantitative methods’ scholarship provided by the University of Leeds.

References

- Adams G, Gulliford M, Ukomunne O, Chinn S, Campbell M (2003) Geographical and organisational variation in the structure of primary care services: implications for study design. *J Health Serv Res Policy* 8(2):87–93
- Appleby J, Thompson J, Jabbal J (2014) How is the health and social care system performing? Quarterly monitoring report July 2014. The King’s Fund, London, http://www.kingsfund.org.uk/sites/files/kf/field/field_publication_file/quarterly-monitoring-report-kingsfund-jun13.pdf. Accessed 25 Oct 2013
- Baltagi B (2006) An alternative derivation of Mundlak’s fixed effects results using system estimation. *Econ Theory* 22(6):1191–1194
- Baltagi B (2008) Econometric analysis of panel data. Wiley, Chichester
- Buckell J, Jones R, Holland D, Batstone G (2013) Efficient thinking: introducing econometric techniques in pathology. *Bull R Coll Pathol* 164:241–243
- Buckell J, Smith A, Longo R, Holland D (2015) Efficiency, heterogeneity and cost function analysis: empirical evidence from pathology services in the National Health Service in England. *Appl Econ* 47(31):3311–3331. doi:[10.1080/00036846.2015.1013617](https://doi.org/10.1080/00036846.2015.1013617)
- Castelli A, Jacobs R, Goddard M, Smith PC (2013) Health, policy and geography: insights from a multi-level modelling approach. *Soc Sci Med* 92:61–73
- D’Amico F, Fernandez J-L (2012) Measuring Inefficiency in long-term care commissioning: evidence from English local authorities. *Appl Econ Perspect Policy* 34(2):275–299

- Department of Health (2006) Report of the review of NHS pathology services in England Chaired by Lord Carter of Coles. Department of Health, London, <http://www.pathologists.org.uk/publications-page/Carter%20Report-The%20Report.pdf> Accessed 25 Oct 2013
- Department of Health (2008) Report of the second phase of the review of NHS pathology services in England Chaired by Lord Carter of Coles. Department of Health, London, http://microtrainees.bham.ac.uk/lib/exe/fetch.php?media=review_report_final_proof08.pdf Accessed 25 Oct 2013
- England NHS (2013) The NHS belongs to the people: a call to action—the technical annex. NHS England, London
- Farsi M, Filippini M, Kuenzle M (2005a) Unobserved heterogeneity in stochastic cost frontier models: an application to Swiss nursing homes. *Appl Econ* 37(18):2127–2141
- Farsi M, Filippini M, Greene W (2005b) Efficiency measurement in network industries: application to the Swiss railway companies. *J Regul Econ* 28(1):69–90
- Felder S, Tauchmann H (2013) Federal state differentials in the efficiency of health production in Germany: an artifact of spatial dependence? *Eur J Health Econ* 14(1):21–39
- Filippini M, Hunt LC (2012) US residential energy demand and energy efficiency: a stochastic demand Frontier approach. *Energy Econ* 34(5):1484–1491
- Greene W (2004) Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health Econ* 13(10):959–980
- Greene W (2008) The econometric approach to efficiency analysis. In: Fried H, Lovell K, Schmidt S (eds) *The measurement of productive efficiency and productivity growth*. Oxford University Press, Oxford
- Greene W (2012a) LIMDEP version 10. Econometric Software, New York
- Greene W (2012b) LIMDEP econometric modelling guide. Econometric Software, New York
- Healthcare Commission (2007) Getting results: pathology services in acute and specialist trusts. Commission for healthcare audit and inspection, London, <http://www.bipsolutions.com/docstore/pdf/16479.pdf>. Accessed 12 Jan 2013
- Holland D, Trigg G (2011) Clinical biochemistry/chemical pathology core report. National Pathology Benchmarking Review 2011/2012. Keele University, Staffordshire
- Hollingsworth B (2008) The measurement of efficiency and productivity of health care delivery. *Health Econ* 17(10):1107–1128
- Hollingsworth B (2012) Revolution, evolution, or status quo? Guidelines for efficiency measurement in health care. *J Prod Anal* 37:1–5
- Jondrow J, Knox Lovell CA, Materov IS, Schmidt P (1982) On the estimation of technical inefficiency in the stochastic frontier production function model. *J Econometrics* 19:233–238
- Kendall MG, Gibbons JD (1990) Rank correlation methods, 5th edn. Oxford University Press, New York
- Kumbhakar SC, Heshmati A (1995) Efficiency measurement in Swedish dairy farms: an application of rotating panel data, 1976–88. *Am J Agric Econ* 77:660–674
- Kumbhakar S, Lovell CAK (2000) Stochastic Frontier analysis. Cambridge University Press, Cambridge
- Kumbhakar S, Lien G, Hardaker JB (2014) Technical efficiency in competing panel data models: a study of Norwegian grain farming. *J Prod Anal* 41:321–337
- Moulton B, Randolph W (1989) Alternative tests of the error components model. *Econometrica* 57:685–693
- Mundlak Y (1978) On the pooling of time series and cross section data. *Econometrica* 46(1):69–85
- Murillo-Zamorano L, Petraglia C (2011) Technical efficiency in primary health care: does quality matter? *Eur J Health Econ* 12(2):115–125
- Mutter R, Greene W, Spector W, Rosko M, Mukamel D (2013) Investigating the impact of endogeneity on inefficiency estimates in the application of stochastic frontier analysis to nursing homes. *J Prod Anal* 39:101–110
- NHS England (2014) Pathology quality assurance review. NHS England, London

- Olsen KR, Street A (2008) The analysis of efficiency among a small number of organisations: how inferences can be improved by exploiting patient-level data. *Health Econ* 17(6):671–681
- Schmidt P, Sickles RC (1984) Production frontiers and panel data. *J Bus Econ Stat* 2(4):367–374
- Smith A, Wheat P (2012) Estimation of cost inefficiency in panel data models with firm specific and sub-company specific effects. *J Prod Anal* 37(1):27–40
- Smith P, Mossialos E, Papanicolas I (2012) Performance measurement for health system improvement: experiences, challenges and prospects. In: Figueras J, McKee M (eds) *Health systems, health, wealth and societal well-being: assessing the case for investing in health systems*. McGraw-Hill, New York, pp 247–280
- Sørensen TH, Olsen KR, Gyrd-Hansen D (2009) Differences in general practice initiated expenditures across Danish local health authorities—a multilevel analysis. *Health Policy* 92(1):35–42
- Wheat P (2014) Econometric cost analysis in vertically separated railways. Doctoral thesis, University of Leeds
- Zhang X, Hauck K, Zhao X (2013) Patient safety in hospitals—a Bayesian analysis of unobservable hospital and speciality level risk factors. *Health Econ* 22(9):1158–1174

Chapter 9

Is There Evidence of ICT Skill Shortages in Canadian Taxfiler Data?

Brian Murphy, Michael R. Veall and Yan Zhang

Abstract Productivity and growth may be affected by what are called “shortages” of specific types of workers. We examine Canadian data for evidence of a shortage of Information and Communication Technology (ICT) workers. Published vacancy and unemployment data is too coarse at the industry level. Accordingly we use two types of administrative data to look for evidence of rising ICT employment and labour income which might indicate a shortage. One dataset is available with little lag in cross section (from payroll records) and the other longitudinal dataset (based on taxfiler data) is available with a 2-year lag. Our results suggest that both data sources may be useful in this instance, with the longitudinal data used to check for compositional changes in the more timely cross section data. Similar approaches may be available for other countries. These data sources provide at most mild evidence of a shortage of Canadian ICT workers in recent times.

Keywords Productivity • Labour markets • Labour income • Vacancies • Unemployment • Longitudinal administrative data

This paper represents the views of the authors and does not reflect the opinions of Statistics Canada. The authors thank a reviewer and Marc Frenette for useful comments, Julien Dicaire for research assistance and the Social Sciences and Humanities Research Council of Canada for financial support.

B. Murphy (✉) • Y. Zhang

Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON K1A 0T6, Canada
e-mail: Brian.Murphy@statcan.gc.ca; Yan.Zhang@statcan.gc.ca

M.R. Veall

Department of Economics, McMaster University, 1280 Main Street West,
Hamilton, ON L8S 4L8, Canada
e-mail: veall@mcmaster.ca

9.1 Introduction

Productivity depends upon the economic efficiency of the labour market. Economic growth could be impeded by “shortages” of certain kinds of workers, particularly workers who may have a key role in innovation, such as science and engineering workers.

Hence information about the existence of such shortages, if any, and their depth and nature, could be an important input into decisions involving productivity, immigration and education policy. Detailed vacancy data would likely be the best information. In its absence, individual longitudinal labour market data on employment and labour income at a detailed industry level may be useful. In this study, we examine this latter possibility using Canadian large-sample, individual, longitudinal taxfiler data coded by industry of employer. We focus on workers associated with the key information and communications technology (ICT) industry. Some of the insights regarding the advantages and drawbacks of such data will extend to the potential use of similar data in other countries.

9.2 Some Relevant Literature

We note at the outset that there is no general agreement on the concept of a labour shortage. Some economists argue that there never can be a shortage. The wage would rise and the market would clear.¹ If one softens that position to allow for a temporary shortage as markets adjust, then wages would rise for as long as the shortage² persisted. Implicitly this analysis is rooted in the supply and demand model which assumes competitive markets.³ There is the possibility of shortage

¹Teitelbaum (2014) considers this view in his book *Falling Behind? Boom, Bust and the Global Race for Scientific Talent*. Indeed he argues that government research policy in the United States has led to something closer to a surplus, which has yet to be cleared by stagnant or falling rates of compensation to science and technology workers.

²Arrow and Capron (1959) define a shortage as “a situation in which there are unfilled vacancies in positions where salaries are the same as those currently being paid to others of the same type and quality”.

³Modifications could be made to the standard supply and demand model which might allow a temporary shortage. For example, if wages are rigid downward, firms may not raise wages now if there is a possibility of a future surplus. Alternatively, higher wages could disproportionately attract high-turnover employees. Other factors besides wages may clear the market. Backes-Gellner and Tuor (2010) discuss German evidence that one method firms use to avoid shortages is to signal that they are good employers, for example by providing ongoing training programs and having a works council. Perhaps these approaches reduce shortages of the attitudinal and motivational skills that Green et al. (1998) find that surveyed employers often emphasize. Healey et al. (2012) find that employers often respond to shortages by lengthening work hours in the short run and by training in the medium run, rather than by raising wages for prospective hires. Richardson (2007) emphasizes more broadly that firms often adjust hiring standards rather than wage offers.

with noncompetitive markets which include models characterized by search and bargaining.⁴

Within Canada, Industry Canada (2010), Information and Communications Technology Council (2011) and Nordicity (2012)⁵ all argued that available evidence pointed to existing or impending shortages of at least some categories of skilled information and communication technology (ICT) workers. Other research has identified relatively low use of ICT as a key factor hampering Canadian productivity growth (e.g. Deloitte 2012).

Therefore the first reason we examine data on ICT labour markets is that they have been used widely as an example of shortage. A second relates to the increased interest in inequality, particularly top-end inequality (Atkinson et al. 2011; Piketty 2014). For Canada, Saez and Veall (2005, 2007), Murphy et al. (2007), Fortin et al. (2012) and Veall (2012) have documented various aspects of a recent surge in top end incomes. Of the many possible reasons for this discussed in the above articles, one is the well-known hypothesis of skill-biased technical change (surveyed, e.g., by Acemoglu 2002). Any evidence of rising ICT wages would be germane to the hypothesis that the surge in top incomes could be associated with the increasing importance of ICT (e.g. Autor and Krueger 1998; Autor et al. 2003). Michaels et al. (2014) study data from 11 countries (not including Canada) and find evidence that there is greater income polarization in ICT industries.

There is also the possibility that the observed phenomenon is not a true shortage but rather a horizontal demand curve for labour determined by the world price of ICT labour (productivity-adjusted). Each firm would like to buy more labour at the current price but is restrained by supply, yet there can be no upward wage pressure. A post by Ozimek (2013), a related post by Cowen (2013) and a presentation by Veall (2013) related to this paper explore this possibility.

⁴These include imperfect competition models where a firm exercises its market power by restraining its own purchases to keep the wage from rising. (A related model would posit that a firm with some market power may restrain its offers to new employees to be no greater than the wages it pays to current employees, because to do otherwise would damage morale and productivity.)

Even with firms that are not large enough to have conventional market power, the no-recall search/bargaining model of Diamond (1971) also implies a monopsonistic wage. The models of Burdett and Mortensen (1998) and Mortensen and Pissarides (1999) posit wage determination mechanisms that can also leave shortages without rising wages.

⁵Subsequently in this paper we focus on certain industry and occupation classification codes. Industry Canada (2010) does not specify its comparable codes but refers more generally to the information and technology industry. Nordicity (2012) uses occupation codes which overlap significantly with ours. They include electronic and electrical engineers; we do not. We also include various types of computer and web technicians; they do not. Information and Communication Technology (2011) uses the intersection of fairly broad occupation codes with industry classifications which are similar to ours. Because they are able to use a somewhat finer industry disaggregation in custom runs on their data, they are able to separate out and hence include a few additional industries which we do not use, principally in computer and communications equipment manufacturing.

9.3 Existing Indicators of the State of the ICT Labour Market

The clearest measure of shortage is probably the vacancy rate. Statistics Canada has only estimated the vacancy rate for Canada since 2011.⁶ The Job Vacancy Statistics (JVS) use the Business Payrolls Survey (BPS) within the Survey of Employment, Payroll and Hours (SEPH). The BPS sample is stratified and consists of 15,000 establishments, with an 80 % response rate and a 30 % imputation rate. The relevant population of establishments is about 900,000. Accordingly, the estimates are by “sector” (2-digit North American Industrial Classification System (NAICS) industry code) and not by occupation and consist of an estimate of the number of vacant positions within an industry divided by the sum of the estimated number of vacant positions and the estimated number of filled positions. It may not be possible to obtain more disaggregated industry estimates with a sample of this size: occasionally estimates are not currently reported for some 2-digit codes because the data are not judged to be of sufficient accuracy. The closest 2-digit NAICS codes to ICT are for Information and cultural industries (NAICS 51)⁷ and Professional, scientific and technical services (NAICS 54).⁸ Between 2011 and 2014II,⁹ the vacancy rate in NAICS 51 rose from 1.5 to 1.8 while remaining stable at 2.2 in NAICS 54. This compares to an aggregate estimate for all classified industries of 1.6 in 2011 and 1.5 in 2014II.¹⁰ This is perhaps weak evidence of a shortage. However, the more important message may be that current job vacancy data for Canada may be at too high a level of industry aggregation to study effectively specific labour markets like that for ICT workers.

Sometimes the vacancy data is combined with unemployment data from the Labour Force Survey (to yield an unemployed-per-available-job estimate). Sometimes industry unemployment data is used on its own to study labour market

⁶It formerly estimated the vacancy rate between 1971 and 1978. Morissette and Zhang (2001) and Galarneau et al. (2001) provided information for 1999. Drummond et al. (2009) emphasized this gap in Canada’s labour market information.

⁷Consisting of Publishing industries (except internet) (511), Motion picture and sound recording industries (512), Broadcasting (except internet) (515), Telecommunications (517), Data processing, hosting and related services (518) and Other information services (519).

⁸Consisting of Legal services (5411), Accounting, tax preparation, bookkeeping and payroll services (5412), Architectural, engineering and related services (5413), Specialized design services (5414), Computer systems design and related services (5415), Management, scientific and technical consulting services (5416), Scientific research and development services (5417), Advertising, public relations and related services (5418) and Other professional, scientific and technical services (5419).

⁹The survey has not been conducted long enough to allow for seasonal adjustment.

¹⁰Another measure in the Bank of Canada’s Business Outlook Survey (2013) comes from the responses to “Does your firm face any shortages of labour that restrict your ability to meet demand?” The percentage of yes answers was around 60 % in 2000, fell almost to 10 % in 2009, rose up to 33 % in 2012III and was 30 % in 2014II. This measure is not disaggregated by industry.

shortages (as in Nordicity 2012). Again, the level of disaggregation is typically to the 2-digit NAICS level. In addition, the Statistics Canada Labour Force Survey designates the industry of unemployment as the last industry of employment (provided such employment occurred within the previous 12 months). The results can sometimes be surprising. For example, the estimated overall unemployment rate in Canada in 2013 was 7.1 %. Information and cultural industries (NAICS 51) is combined with Arts, entertainment and recreation (NAICS 71) in these tables and had an estimated unemployment rate of 5.9 % while Professional, scientific and technical services (NAICS 54) had an unemployment rate of 3.5 % which some would say was clear evidence of tightness. But note that Educational services (NAICS 61) also had a low unemployment rate of 3.8 % despite such a slack labour market for teachers that a significant number choose to leave Canada (e.g. Sagan 2013). It would seem very likely that the low unemployment rate in Educational services is because people who lose their jobs as teachers do not wait for new full-time vacancies to appear in such a poor labour market but either take part-time work as supply teachers or switch to other industries. (The number of unemployed in Educational Services also does not include the large number of individuals who have trained as teachers but have never been employed as teachers (Sagan 2013).) Just as low unemployment rates do not always indicate tight labour markets, higher unemployment rates, such as the estimated 7.8 % in Construction (NAICS 23), may sometimes be due to workers being unemployed for short periods while waiting for new jobs, frequently in the same industry.

There are also estimates of the unemployment rate by occupation. Using the National Occupation Code (NOC) classification known as NOC-S, the 2013 unemployment rate for Natural and applied sciences and related occupations (NOC-S code C, the closest NOC-S code for ICT workers) was 3.2 %, well below the overall rate of 7.1 %. But again the Labour Force Survey only counts someone as being unemployed in an occupation if they have had a position in that occupation within the previous 12 months. Hence the unemployment rate associated with a particular occupation will not include individuals who were in that occupation, lose their jobs and switch to another occupation, perhaps one where prospects will be better. Similarly an occupation's unemployment rate will not count those who are unemployed because they cannot find work in an occupation they have trained for. Again the illustration is from education, where the 2013 unemployment rate for Teachers and professors (NOC-S E1 and E130, combined in the data) is a low 3.3 %, again inconsistent with the poor labour market for teachers described in Sagan (2013).

Our conclusion is that 2-digit NAICS code classifications are too coarse to examine the ICT industries using either Job Vacancy Statistics or other measures. Unemployment rates by industry or occupation have the additional problem in examining labour shortages that they miss individuals who become unemployed in an industry/occupation but find work in another industry/occupation (perhaps because prospects in the initial industry/occupation are poor) and individuals who have trained for an industry/occupation but never find employment in it.

Given the current size of the relevant data collection surveys in Canada, vacancy or unemployment data is unlikely to be able to identify labour market shortages in specific industries such as ICT. There is, however, the possibility of using administrative data on employment and labour incomes to identify conventional symptoms of a labour shortage (i.e. either rising employment, rising wages, or both).

Let us first turn to employment and labour income measures that originate in the Statistics Canada Survey of Employment, Payroll and Hours (SEPH) mentioned above. While it has a survey element as its name suggests, the employment and labour income measures are produced from administrative payroll data submitted to the Canada Revenue Agency for the population of 900,000 establishments. These data are classified by NAICS industry, typically at a significantly more detailed level than the Labour Force Survey provides. In our case, the data are available for NAICS 517 (Telecommunications) and NAICS 518 (Data processing, hosting and related services: equivalent to the 6-digit code, 518210) which are much better matches with ICT than the 2-digit categories discussed above. The current SEPH series for employment and average weekly earnings begins in 1991. However, administrative data were only introduced in 2001, which affects the comparability of pre- and post-2001 estimates.

Figure 9.1 shows that employment in Data processing, hosting and related services (NAICS 518) has increased much more quickly than in Information and cultural services (NAICS 51), emphasizing that the latter is not strongly correlated with the former. But employment in Telecommunications (NAICS 517) declined. Professional, scientific and technical services (NAICS 54) employment has also increased rapidly relative to the overall trend in employment for the industrial aggregate excluding unclassified businesses.

In Fig. 9.2, the SEPH data are used to study average weekly earnings. All the categories have recent fairly modest rates of wage increase except for Data processing, hosting and related services (NAICS 518) which had much larger rates of increase until plunging in 2013. There has only been a slight upward trend in NAICS 517, but a noticeable upward movement in 2013.

Taken together, the SEPH information seems to indicate increased employment in NAICS 518 but a recent decline in compensation and no increase (actually a decline) in employment in NAICS 517 with a recent increase in compensation. Hence there was no clear indication of shortage as of 2013.

There are two main shortcomings to using SEPH data to examine the labour market for ICT workers. First, firm composition within the ICT industry changes over time as firms are founded, fail, expand or contract. This can lead to changes in measured employment and labour income that have little relationship to overall industry conditions. Second, industry data will not match occupation. Within each industry there are many different occupations (whose mix will change over time). For example if firms begin to contract out accounting that formerly was done internally, this will change the weekly earnings measures. The next section will address how using the longitudinal feature of Canadian taxfiler data may address these shortcomings.

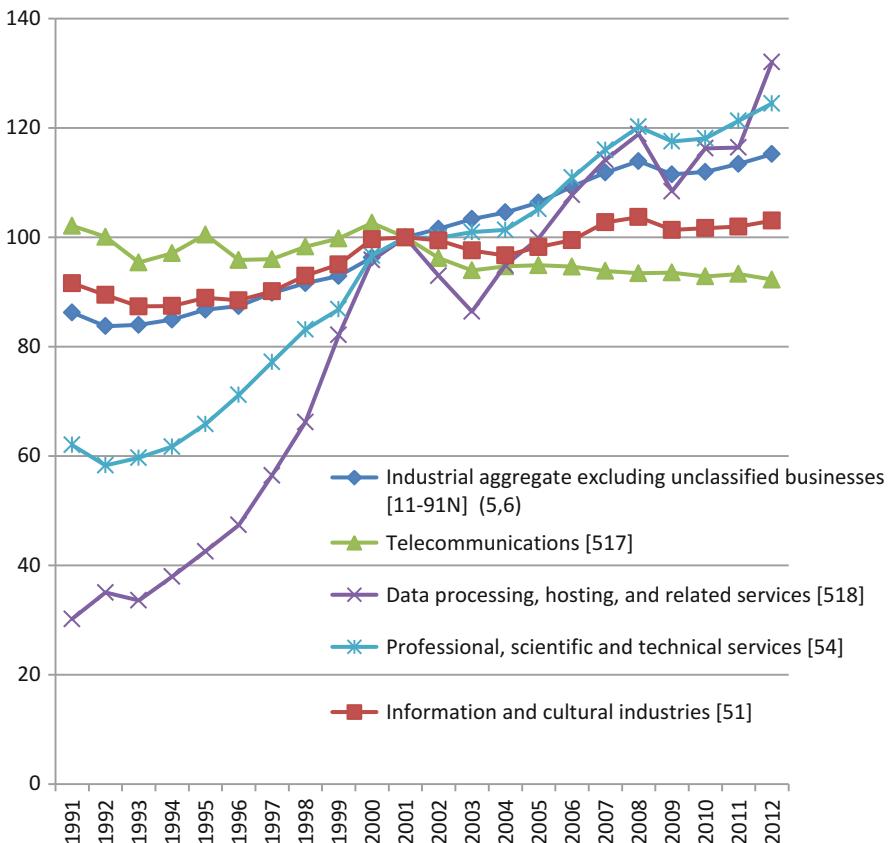


Fig. 9.1 Employment, categorized utilizing NAICS classification (2001 = 100). *Source:* Statistics Canada, CANSIM 281-0024 Employment (Survey of Employment, Payroll and Hours), by type of employee for selected industries classified using the North American Industry Classification System (NAICS), annual

9.4 Using Taxfiler Data to Analyze ICT Employment and Wage Increases

We explore the use of the taxfiler data available in the Canadian Longitudinal Administrative Databank (LAD). The LAD is an anonymized, annual 20 % sample of taxfilers for Canada from 1982 to 2010. It most recently contains about five million tax records per year. We restrict our sample to workers.¹¹ Records are linked longitudinally.

¹¹Workers are taxfilers who had employment income reported on slips for personal income tax purposes. The sample size is about three million per year from 2005 to 2010.

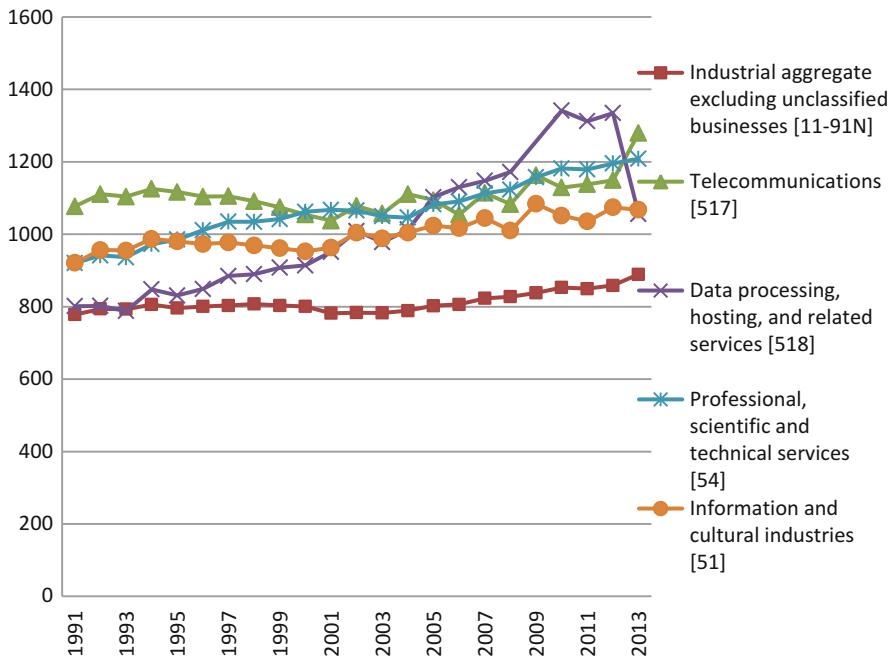


Fig. 9.2 Average weekly earnings by NAICS classification, \$2010. Source: Statistics Canada, CANSIM Table 281-0027 Average weekly earnings (Survey of Employment, Payroll and Hours), by type of employee for selected industries classified using the North American Industry Classification System (NAICS), annual (\$2010). The 2009 observation for NAICS 518 was not released because Statistics Canada judged it to be of insufficient accuracy: the graph has been completed by interpolation

While there is the clear disadvantage that the latest LAD data available is for 2011, there is the clear advantage that the longitudinal aspect allows the comparison of the *same* individuals over time, mitigating issues associated with the changing composition of any classification. The LAD has 3-digit NAICS coding,^{12 13} since 2000, although unfortunately if understandably, the NAICS codes change 3 times

¹²The NAICS codes on the LAD are derived from a linkage of employment income tax files and Statistics Canada's business register. In cases where a tax filer worked for more than one employer (i.e. more than one T4) in a calendar year, his/her NAICS code refers to the industry of the job on which he/she made the highest T4 income.

¹³A small proportion of our sample's NAICS codes were missing, which account for 1.36–4.12 % of the sample, varying by years.

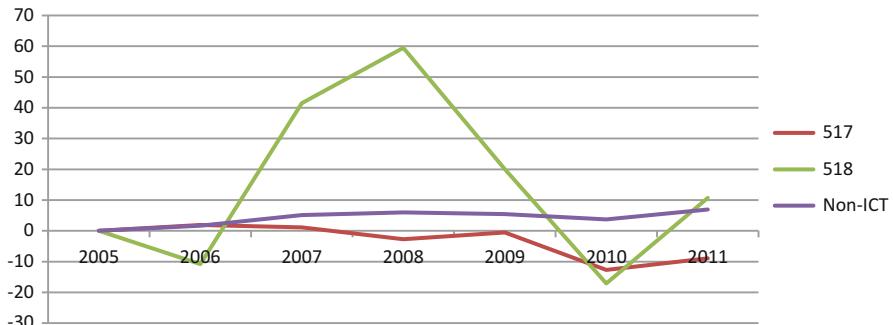


Fig. 9.3 # of labour income recipients (tax files), % change from 2005, 25–64. *Source:* Statistics Canada, special tabulation on the Longitudinal Analysis Databank. *Note:* In this and subsequent Figures, “Non-ICT” means individuals for whom there is a NAICS code, but it is neither NAICS 517 nor NAICS 518

during the sample.¹⁴ The 1997 classification is used for 2000 and 2001, the 2002 classification is used for 2002–2004 and the 2007 classification is used for 2005–2011. Hence we will focus on the 2005–2011 period.

Figure 9.3 gives the percentage change from 2005 in the number of wage earners (those with positive entries on their T4 slips which report labour income for tax purposes) for our two key industries. For NAICS 517, there was a steady, slow decline, similar to Fig. 9.1 from the SEPH data. For NAICS 518, there was a boom in 2007 and 2008 followed by a bust in 2009 and 2010, with a rebound in 2011.

Figure 9.4 gives average T4 income for those age 25–64. (The graph for all ages is identical in pattern with somewhat lower levels.) The stable pattern for 517 is similar to that for the SEPH in Fig. 9.2 up to 2011. For 518, recall that the SEPH observation for 2009 is not available, but that there was a sizeable 2008–2011 increase. That pattern is not visible in these LAD data.

In Fig. 9.5 we move to a cohort approach where we examine the incomes of three cohorts all age 25–64 in 2005: (1) those who worked in NAICS 517 in 2005 (2) those who worked in NAICS 518 in 2005, and (3) those who worked in an industry other than NAICS 517 or 518 (what we call Non-ICT) in 2005. Following cohorts avoids issues of changing industry composition (although differences in trend may be due to many factors besides different levels of labour “shortage”). It also includes those who switch industries of employment or who are unemployed longer than 12 months, unlike the Labour Force Survey measures. In any case, the income level for

¹⁴NAICS are revised every 5 years to keep the classification system current with changes in economic activities since its inception in 1997. Therefore the components of the same NAICS code could vary from one version of NAICS to the other. For example, in NAICS 2002, 518 include Internet Service Providers, Web Search Portals, and Data Processing Services. After the revision of 2007, NAICS 518 only include Data Processing, Hosting, and Related Services, while Internet service providers became part of 517, and web search portals became part of 519.

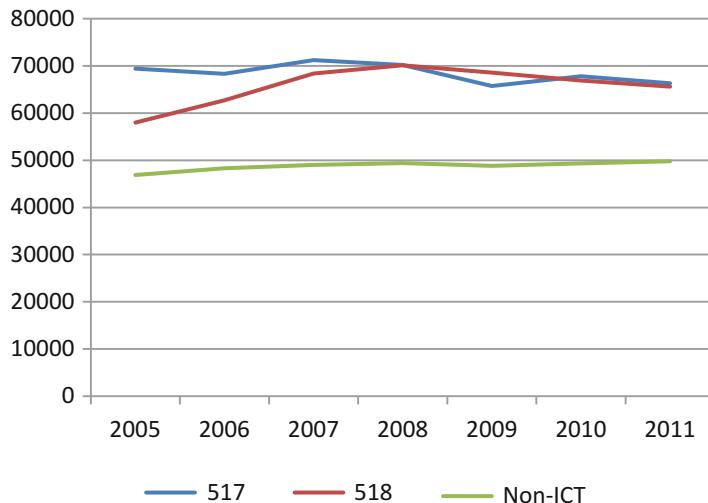


Fig. 9.4 Average employment income per recipients (tax files), Age 25–64, \$2010. *Source:* Statistics Canada, Special tabulation on the Longitudinal Analysis Databank

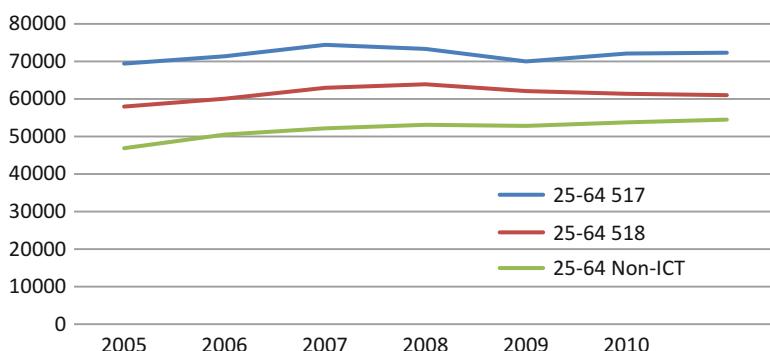


Fig. 9.5 Average employment income, (\$2010) cohort, Age 25–64 and Working in NAICS Industry in 2005. *Source:* Statistics Canada, special tabulation on the Longitudinal Analysis Databank

NAICS 517 is higher in Fig. 9.5. For NAICS 518 there is less run-up from 2005 to 2008 but there is the same mild decline in the last 2 years.

Figure 9.6 repeats this type of cohort analysis, restricting the cohorts to those aged 25–44 in 2005. The levels are somewhat lower than in Fig. 9.5 and there is more sign of a trend increase. There is a hint of an uptick for NAICS 517 in 2010 and 2011. For NAICS 518, the slight 2009 and 2010 downturn in Fig. 9.5 corresponds to a plateau in Fig. 9.6. But combined with Figs. 9.3, 9.4 and 9.5, our overall conclusion is that the LAD data would have provided at most mild evidence of a labour shortage in the ICT industries in 2011.

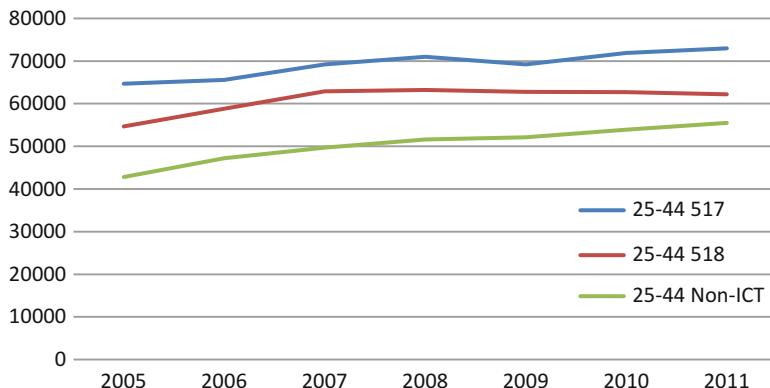


Fig. 9.6 Average employment income, (\$2010) cohort Age 25–44 and Working in NAICS Industry in 2005. *Source:* Statistics Canada, special tabulation on the Longitudinal Analysis Databank

We also can use the LAD data to examine the income experience of recent immigrants, who are identified in the data. While the LAD does not contain information on occupation in general, it does include intended occupation upon landing for immigrants¹⁵ who landed 1980 or thereafter. Occupations are coded by the National Occupation Code (NOC; 2006 classification). We describe ICT immigrants as those whose intended occupation upon landing was in one of the NOC codes Computer operators (1421), Computer engineers (2147), Computer systems analysts (2162), Computer programmers (2163), Information systems analysts and consultants (2171), Database analysts and data administrators (2172), Software engineers (2173), Computer programmers and interactive media developers (2174), Computer and network operators and web technicians (2281), User support technicians (2282) and Systems testing technicians (2283).

For Fig. 9.7, as a first step, we do not use the occupation codes but follow a cohort of immigrants who in 2005 were working in NAICS 517 as compared to those working that year in what we are calling non-ICT industries, that is neither NAICS 517 nor NAICS 518. (The numbers are too small to do this analysis for NAICS 518.) The Fig. 9.7 shows that the 517 immigrant cohort has received higher average employment income than the non-ICT immigrant cohort, and that the difference is larger for those who recently landed. For those who landed from 1980 to 1994 (some as children), incomes rise slowly and steadily between 2005 and 2011. For those who landed between 1995 and 1999, there is a modest increase between 2009 and 2010. The most recently landed who worked in NAICS 517 in 2005 seem to have had bumpier incomes with a dip in 2009, a jump in 2010 and a dip in 2011. Hence the evidence suggests that immigrants who worked in NAICS 517 in 2005

¹⁵More precisely, it is only available for the principal claimants, often regarded as the heads of households.

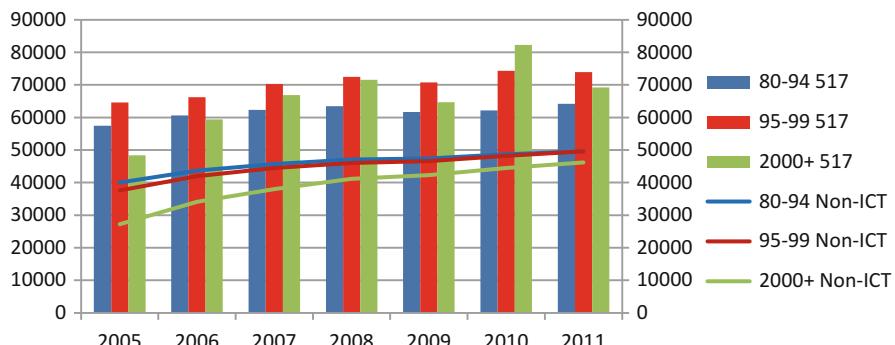


Fig. 9.7 Average employment income (\$2010) of immigrants by landing year, cohort working in NAICS 517 in 2005, Age 25–44. *Source:* Statistics Canada, special tabulation on the Longitudinal Analysis Databank

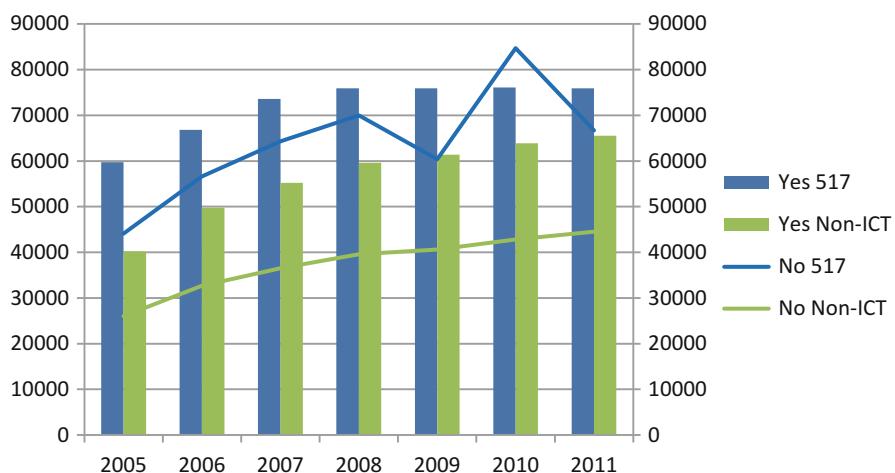


Fig. 9.8 Average T4 income (\$2010) of immigrants by occupation intention, cohort working in NAICS 517 in 2005, Age 25–44, Landed in 2000 and after. *Source:* Statistics Canada, special tabulation on the Longitudinal Analysis Databank

had relatively high incomes between 2005 and 2011, but it is less clear that there has been a relative increase over that period.

Figure 9.8 introduces our use of the immigrant data for intended occupation upon landing. It compares four cohorts of recent immigrants, all of whom were age 25–44 in 2005. The Yes 517 cohort is so-named because yes, it had an intended ICT occupation (as described earlier) upon landing and was working in 2005 in NAICS 517, where again the numbers in NAICS 518 are too small for this analysis. The Yes Non-CIT cohort had an intended ICT occupation upon landing but was not working in 2005 in NAICS 517 or 518. The No 517 cohort did not have an intended ICT occupation but was working in 2005 in NAICS 517. The No Non-ICT cohort did

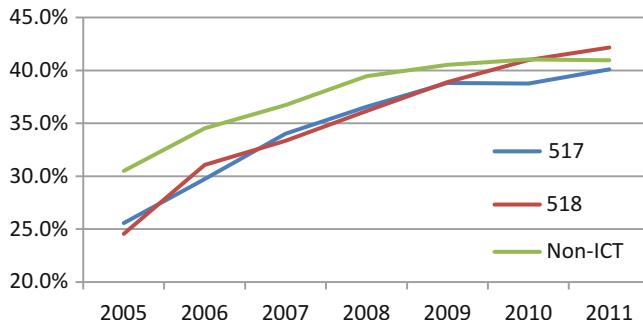


Fig. 9.9 New (2000+) immigrants as % all immigrants (1980+) within sector

not have an intended ICT occupation nor was it working in 2005 in NAICS 517 or 518. This last cohort with no ICT connection had significantly lower incomes from 2005 to 2011. For the three cohorts with some ICT connection, there appears to be little evidence of unusual income increases in the 2008–2011 period that would correspond to a shortage although the No 517 cohort average income falls in 2009, spikes in 2010 and then falls again in 2011.

Figure 9.9 shows that the share of recent immigrants working in an industry as a percentage of all immigrants working in that industry is rising faster in 517 and 518 than in other industries. While evidence of the industry drawing new immigrants, even with the increase the shares in 517 and 518 have only attained the levels of other industries.

It is possible that differences in language or credential/experience recognition make the time path of the incomes of immigrant ICT workers uninformative about the time paths of other ICT workers. Nevertheless, the evidence of Figs. 9.7, 9.8 and 9.9 together suggest that while immigrants who likely have ICT skills have had relatively high incomes, and ICT industries have increased their share of recent immigrants, their experience as reported in the LAD provides only a little evidence of an effect on them of any ICT skills shortage as of 2011.

9.5 Discussion and Conclusions

Even if one puts aside the conceptual issues in considering a labour shortage, there are measurement issues. In Canada the surveys that provide estimates of vacancies, unemployment and other labour force aggregates do not publish sufficient disaggregation to provide the detailed information required to identify shortages in specific industries or occupations, nor do they likely have sufficient sample to do so. There can also be difficulties in unemployment-based measures of labour market shortage by industry or occupation, because individuals can change industry or occupation. Finally industry-based measures are subject to compositional changes,

for example with firms changing which activities they do inhouse, changing their main activities or merging. We illustrate these problems by trying to use the existing survey data sources to provide evidence regarding possible shortage of Information and Communication Technology (ICT) workers.

We continue by exploring the alternative of using large-sample administrative data that permits better disaggregation by industry. While vacancies are not observable, potential industry labour shortages could be identified by industry increases in employment and/or labour income. One data set we use is the cross sectional Survey of Employment, Payroll and Hours (SEPH), which has an administrative payroll data backbone sufficient to provide estimates of employment and labour income at the 3-digit level in the North American Industrial Classification System (NAICS). For the ICT, industries, we use two NAICS codes: NAICS 517 (Telecommunications) and the smaller NAICS 518 (Data processing, hosting and related services). There is no information regarding occupation.

Another possible data set is the Statistics Canada Longitudinal Database (LAD, based on taxfiler data) which we use to provide a cohort analysis for NAICS 517 and 518, minimizing the problems associated with job switchers and changing industry composition. This data set is only currently available with a 2-year lag. While occupation information is not generally available, there is intended occupation upon landing for principal-applicant immigrants since 1980. Hence we also follow cohorts of immigrants whose intended occupation were in ICT areas.

As of 2011, our SEPH cross section and LAD cohort approaches find only weak evidence of labour shortages in NAICS 517 or NAICS 518. We also find no strong evidence of such shortages in the more recent SEPH data. As an aside, that we do not find evidence of a consistent surge in incomes of ICT workers is not supportive of the theory that the recent rise in top-end inequality is largely “skill-biased” technical change, although it should be emphasized that this is just one fragment of evidence.

In the absence of large-sample vacancy data and unemployment survey data, administrative data on employment and labour income may be useful in examining labour market tightness by industry. In countries such as Canada where longitudinal data are available but on a less timely basis than available cross section data, an approach combining those two types of data sources may be the most effective. The principal data source would be the more timely cross sections with cohorts constructed from the less current longitudinal data used to check for the effects of changes in industry composition.

References

- Acemoglu D (2002) Technical change, inequality and the labor market. *J Econ Lit* 40:7–72
Arrow K, Capron W (1959) Dynamic shortages and price rises: the engineer-scientist case. *Q J Econ* 73:292–308
Atkinson AB, Piketty T, Saez E (2011) Top incomes in the long run of history. *J Econ Lit* 49:3–71
Autor D, Krueger A (1998) Computing inequality: have computers changed the labor market? *Q J Econ* 113:1160–1213

- Autor D, Levy F, Murnane R (2003) The skill content of recent technological change: an empirical exploration. *Q J Econ* 118:1279–1333
- Backes-Gellner U, Tuor SN (2010) On the importance of good work climate and labor relations. *Ind Labor Relat Rev* 63:271–285
- Bank of Canada (2013) Business outlook survey. Bank of Canada, Ottawa
- Burdett K, Mortensen DT (1998) Wage differentials, employer size, and unemployment. *Int Econ Rev* 39:257–273
- Cowen T (2013) Is there a shortage of STEM workers in the United States? <http://marginalrevolution.com/marginalrevolution/2013/04/is-there-a-shortage-of-stem-workers-in-the-united-states.html>. Accessed 21 May 2015
- Deloitte (2012) The future of productivity in Canada, clear choices for a competitive Canada. http://www.deloitte.com/view/en_CA/ca/pressroom/ca-pressreleases-en/a072cc796ba1a310VgnVCM1000003256f70aRCRD.htm. Accessed 1 Mar 2015
- Diamond P (1971) A model of price adjustment. *J Econ Theory* 3:156–168
- Drummond D, Beale E, Kolby K, Loiselle M, Miner R. (2009) Working together to build a better labour market information system for Canada. http://publications.gc.ca/collections/collection_2011/rhdec-hrsdc/HS18-24-2009-eng.pdf. Accessed 21 May 2015
- Fortin N, Green D, Lemieux T et al (2012) Canadian inequality: recent developments and policy options. *Can Public Policy* 38:121–145
- Galarneau D, Krebs H, Morissette R et al (2001) The quest for workers: a new portrait of job vacancies in Canada. <http://www.statcan.gc.ca/pub/71-584-m/71-584-m2001002-eng.pdf>. Accessed 10 May 2015
- Green F, Machin S, Wilkinson D (1998) The meaning and determinants of skills shortages. *Oxf Bull Econ Stat* 60:165–187
- Healey J, Mavromaras K, Sloane PJ (2012) Skill shortages: prevalence, causes, remedies and consequences for Australian businesses. National Vocational Education and Training Research and Evaluation Program, Canberra
- Industry Canada (2010) Improving Canada's digital advantage: strategies for sustainable prosperity. http://publications.gc.ca/collections/collection_2010/ic/lu4-144-2010-eng.pdf. Accessed 1 Mar 2015
- Information and Technology Council (2011) Outlook for human resources in the ICT labour Market, 2011–2016. http://www.ictc-ctic.ca/wp-content/uploads/2012/06/ICTC_Outlook2011_EN_11-11.pdf. Accessed 1 Mar 2015
- Michaels G, Ashwini N, Van Reenen J (2014) Has ICT polarized skill demand? Evidence from eleven countries over 25 years. *Rev Econ Stat* 96:60–77
- Morissette R, Zhang X (2001) Which firms have high vacancy rates in Canada, Statistics Canada. <http://www.statcan.gc.ca/pub/11-015-x/labour-maind/4070301-eng.htm>. Accessed 10 May 2015
- Mortensen D, Pissarides C (1999) New developments in models of search in the labor market. In: Card D, Ashenfelter O (eds) *Handbook of labor economics*. Elsevier, Amsterdam
- Murphy B, Roberts P, Wolfson M (2007) High income Canadians. *Perspect Labour Income* 8:1–13, <http://www.statcan.gc.ca/pub/75-001-x/2007109/article/10350-eng.pdf>. Accessed 25 May 2015
- Nordicity (2012) Labour supply/demand dynamics of Canada's information and technology (ICT) sector. <http://www.nordicity.com/media/20121112pnzutcbz.pdf>. Accessed 11 Mar 2015
- Ozimek A (2013) An alternative theory of the skills shortage. <http://www.forbes.com/sites/modeledbehavior/2013/04/24/an-alternative-theory-of-the-skills-shortage/>. Accessed 21 May 2015
- Piketty T (2014) Capital in the twenty-first century. Harvard University Press, Cambridge
- Richardson S (2007) What is a skill shortage? National Vocational Education and Training Research and Evaluation Program, Canberra. <http://files.eric.ed.gov/fulltext/ED495918.pdf>. Accessed 21 May 2015
- Saez E, Veall MR (2005) The evolution of high incomes in Northern America: lessons from Canadian evidence. *Am Econ Rev* 95:831–849

- Saez E, Veall MR (2007) The evolution of high incomes in Canada. In: Atkinson AB, Piketty T (eds) *Top incomes over the twentieth century: a contrast between continental European and English-speaking countries*. Oxford University Press, Oxford, pp 1920–2000
- Sagan A (2013) New Canadian teachers head abroad amid tight job market. <http://www.cbc.ca/news/canada/new-canadian-teachers-head-abroad-amid-tight-job-market-1.2426110>. Accessed 21 May 2015
- Teitelbaum MS (2014) Falling behind? Boom, bust and the global race for scientific talent. Princeton University Press, Princeton
- Veall MR (2012) Top income shares in Canada: recent trends and policy implications. *Can J Econ* 45:1247–1272
- Veall MR (2013) Labour ‘shortages’ in a globalized ICT industry. Presentation to Spring Policy Conference, Ottawa Economics Association, 27 March 2013

Chapter 10

Worker Separations and Industry Instability

Kim P. Huynh, Yuri Ostrovsky, Robert J. Petrunia, and Marcel-Cristian Voia

Abstract This paper looks at the impact industry instability has on worker separations. Workers leave firms one of two ways: (i) voluntarily by quitting; or (ii) involuntarily through firm layoffs. Using data drawn from the Longitudinal Worker File, a Canadian firm-worker matched employment database, we are able distinguish between voluntary and involuntary separations using information on reasons for separations and assess the impact industry shutdown rates have on worker separation rates, both voluntarily and involuntarily. Once controlling for various factors and potential selection bias, we find that industry shutdown rates have a positive and significant effect on the overall separation, layoff and quit rates of workers. Finally, industry instability has a much larger impact on layoff rates when comparing voluntary and involuntary separations.

Keywords Worker separations • Firm survival • Selection

JEL Classification: J24, J31, J63, C35

K.P. Huynh
Bank of Canada, 234 Wellington Street, Ottawa, ON, Canada K1A 0G9
e-mail: kim@huynh.tv

Y. Ostrovsky
Statistics Canada, 24-J RHC, 100 Tunney's Pasture Driveway,
Ottawa, ON, Canada K1A 0T6
e-mail: yuri.ostrovsky@statcan.gc.ca

R.J. Petrunia
Lakehead University, 955 Oliver Road, Thunder Bay, ON, Canada P7B 5E1
e-mail: rpetrungi@lakeheadu.ca

M.-C. Voia (✉)
Department of Economics, Carleton University, 1125 Colonel By Drive,
Ottawa, ON, Canada K1S 5B6
e-mail: marcel.voia@carleton.ca

10.1 Introduction

Worker separations and potential job instability have wide ranging financial and other consequences for individuals and their families.¹ There is an extensive literature demonstrating the effects of employees human capital, firm characteristics and labour market conditions on job instability. Industry turnover, firm growth and survival, and the nature of competition creates turbulence for workers due to potential job loss.² In this paper, we investigate individual worker instability in the context of overall industry instability by quantifying the impact of firm shutdown rates within an industry on worker separation rates.

A study by Quintin and Stevens (2005) suggests that higher industry exit rates increase worker turnover due to higher worker separation rates in surviving firms. Workers leave firms one of two ways: (i) voluntarily by quitting; or (ii) involuntarily through firm layoffs. We build on Quintin and Stevens (2005) by considering the possibility that high firm turnover leads to higher overall job instability through increases in both voluntary separations and involuntary separations. Layoffs likely are higher in unstable industries with higher firm shutdown rates. There are a number of reasons for workers to quit including moving to a new job, returning to school, temporary parental and pregnancy leaves or retirement. Moving to a new job or returning to school provide two options for a worker to improve their employment arrangement. Returning to school allows a worker to earn higher wages by becoming more productive through the acquisition of human capital and improved skills. High firm instability within an industry likely increases the incentive for workers to quit for these reasons as the increase in uncertainty lowers the expected benefits of current employment. Further, on the job learning by doing allows the worker to acquire job specific skills and human capital. Kambourov and Manovskii (2009) argue that the benefits of these job specific skills fall if employment spells are short. Thus, workers may choose to quit rather than accumulate skills through learning by doing in anticipation of short employment. Industry instability creates instability for firms, workers and the nature of firm-worker relationships.

The novelty of the study is that we are able to distinguish between voluntary and involuntary separations using information on reasons for separations provided by employers. We use a rich administrative Canadian employer-employee dataset called the Longitudinal Worker File (LWF). The database contains information on the reasons for separation, quit versus layoff. In Canada, employers are by law required to provide such information. The longitudinal nature of LWF allows us

¹For example, Jacobson et al. (1993), Gottschalk and Moffitt (1994), Gottschalk and Moffitt (2009), Beach et al. (2003) and Morissette and Ostrovsky (2005) study impacts on worker wages; Browning and Lusardi (1996) investigates consumption decisions; Pistaferra (2003) shows altering family savings and labour supply decisions; Guiso et al. (2002) looks at occupational choices; and Fraser (2001) even demonstrates impacts on fertility behaviour.

²Huynh and Petrunia (2010) and Huynh et al. (2011) investigate firm dynamics and industry instability.

to distinguish between different cohort of workers and control for their unobserved characteristics such as age, tenure, place of residence, marital status and sex.

We estimate a probit and bivariate probit with selection to gauge the effects of industry shutdown rates on the probability of worker separations. Once controlling for observables and potential selection bias, the results of our study are:

1. Industry shutdown rates have a positive and statistically significant effect on the probability of worker separations, firm layoff rates and worker quit rates. From the worker's view, separation involves both voluntary quits and involuntary layoffs. Industry instability increases both voluntary and involuntary separation of workers.
2. Quantitatively, industry shutdown rates have a much bigger impact on firm layoff rates than worker quit rates. A 1 % increase in the industry shutdown rate causes a 0.676 increase in the firm layoff rate, but only a 0.023 increase in worker quit rates. Thus, most of the impact industry instability has on worker separation is involuntary from the worker's perspective.
3. Union membership lowers the impact of industry instability on worker separation, layoff and quit rates.

The results are particularly interesting since we are able to isolate the effects of worker characteristics, firm characteristics and labour market conditions. Our data allow us to consider these effects simultaneously as they include information about firms (size, payroll) as well as individual characteristics (age, tenure, place of residence, etc.). Huynh et al. (2014) also look at the impact of industry instability, but focus instead on permanent layoffs and wages as measures of worker outcomes. This study shows that industry instability also affects worker turnover through workers quitting to seek alternative opportunities.

A firm's shutdown implies a separation for all the workers at the firm. A high industry shutdown rate leads to high worker separation rate because of individual firm shutdown. However, our empirical analysis focuses on worker separations at continuing firms. Thus, the positive impact of industry instability on worker separations does not directly result from worker separations at individual firms shutting down.

The rest of the paper is organized in the following fashion: the LWF is described in Sect. 10.2 while the empirical methodology is discussed in Sect. 10.3. The results are discussed in Sect. 10.4 and finally Sect. 10.5 concludes.

10.2 The Longitudinal Worker File and EUKLEMS Data

The data are from the Longitudinal Worker File (LWF) administrative datafile for the period spanning 1992–2004. The LWF contains annual information on a 10 % random sample of all tax filers. We keep individuals living in the 10 Canadian provinces who are between 25 and 64 years of age. The LWF links information from four sources:

1. The T4 Supplementary Tax File—the source is data on individuals taken from their T4 statement of remuneration forms. Employers issue a T4 form to any employee with earnings above a minimum triggering payroll deductions such as income tax, employment insurance premiums and Canada pension plan premiums. Each T4 form provides an individual's annual earnings from a given employer along with taxes and other payroll deductions, pension contributions and union dues;
2. The Record of Employment (ROE)—Canadian law requires employers to issue a ROE for any employee separation. The ROE provides the reason and type of separation. Voluntary reasons for separations include leave of absence, injury or illness, quit, pregnancy and parental leaves, and retirement, while involuntary reasons for separations include shortage of work, labor dispute, and firing. Both involuntary and voluntary reasons can result in a permanent or temporary separation;
3. Longitudinal Employment Analysis Program (LEAP) database—LEAP is an firm level database, which includes all firms with a positive payroll. LEAP provides the size of the employee's firm, but also allows the tracking of worker movement across firms;
4. T1 files—the T1 form is the personal income tax form. This form includes demographic information on individual tax filers such as age, sex, marital status and location.

Firm information from LEAP allows the calculation of industry shutdown rates. A continuing firm has a positive annual payroll, while a zero annual payroll indicates firm shutdown. We restrict our attention to shutdown as exit is harder to define.³ Industry j 's annual specific shutdown rate is the ratio, where the numerator is the total number of firms in industry j with a positive payroll in year t and zero payroll in period $t + 1$, and the denominator is the total number of all firms with a positive payroll in industry j in period t .

Summary statistics for our sample, by industry, is given in Table 10.1. The separation variable is binary with a value of 1 when there is a worker separation from an employer, and 0 otherwise. The separation rate varies from a low of 20 % in Coal and petroleum products, and Chemical and chemical products industries to a high of 51 % in Construction.

10.3 Empirical Methodology

The analysis uses industry shutdown rate to capture industry instability. A firm is said to have shutdown in period t when it has positive payroll in year t and a zero payroll in year $t + 1$. Possible reasons for a firm shutting down include permanent

³All exiting firms shutdown, but not all shutdown are exits. Multiple consecutive years of zero payroll would be needed to identify firm exit. Huynh et al. (2010) investigate firm exit in the context of industry instability.

Table 10.1 Summary statistics

Industries	NAICS	Age	Female	Tenure	Earnings	Separation Rate	# of Firms	# of Jobs
Mining and quarrying	21	41.5	0.2	6.41	57,360	0.31	2,330	16,355
Food products, beverages and tobacco	311–312	40.7	0.4	6.13	31,290	0.34	3,245	29,790
Textiles, textile products, leather and footwear	313–316	41.7	0.6	5.56	21,110	0.37	3,045	14,855
Wood and products of wood and cork	321	40.6	0.1	6.63	38,460	0.34	2,195	14,725
Petroleum and coal products	324	42.0	0.3	8.78	61,450	0.20	85	2,320
Chemical and chemical products	325	40.7	0.4	6.06	48,760	0.20	965	8,765
Rubber and plastics products	326	39.5	0.3	6.00	35,280	0.27	1,430	12,015
Basic metals and fabricated metal products	331–332	41.6	0.2	7.02	41,990	0.28	4,805	25,505
Machinery, nec	333, 3352	40.4	0.2	5.82	42,290	0.26	2,775	13,285
Electrical and optical equipment	334, 3351, 3353, 3359	40.0	0.3	6.47	52,070	0.22	1,595	17,040
Transport equipment	336	41.2	0.2	8.14	51,560	0.34	1,160	24,150
Construction	23	40.7	0.1	4.23	25,070	0.51	35,620	64,580
Sale, maintenance and repair of motor vehicles	415, 441, 447, 8111	39.9	0.3	5.27	29,160	0.24	17,555	34,180
Retail trade, except of motor vehicles	44–45, 8112–8114	39.8	0.6	5.02	19,830	0.26	35,655	106,690
Hotels and restaurants	72	38.2	0.6	3.48	11,330	0.31	31,195	66,140
Transport and storage	48, 493, 5615	41.7	0.3	5.91	32,350	0.27	15,100	53,930
Real estate activities	531	42.7	0.5	4.77	26,600	0.22	7,120	15,950

Note: Age and Tenure are measured in years while Gender is equal to zero if male and one if female. Earnings are measured in Canadian dollars deflated to the Consumer Price Index. The quit rates are in proportions

firm exit and temporary situations due to strikes, lockouts, renovation and capital upgrades, and slumps in demand. Given the nature of the dataset, shutdown is at least year-long closure by a firm. With a year plus prospect of no job, workers are unlikely to wait to return to unstable firm. Therefore, industry shutdown should impact worker separations both directly and indirectly. Direct effect occurs since the rate of layoffs within an industry should increase with an increase in industry shutdown rates as firms shed workers. The indirect effect results from workers quitting in anticipation of a prolonged separation from their firm in the face of higher industry shutdown rates. Thus, the direct effect is an involuntary separation from the worker's point of view, while the indirect effect is voluntary.

Our benchmark model is a reduced form probit model of separations in which the latent dependent variable defined by

$$\text{Worker Sep}_{ijkt}^* = \alpha + \beta X_{jt} + \gamma B_{it} + \sum_{k=1}^K \varphi_k C_{kt} + \sum_{j=1}^J \psi_j I_{jt} + \delta D_t + u_{ijkt}, \quad (10.1)$$

where X_{jt} is the annual shutdown rate in industry j in period t . $\text{WorkerSep}_{ijkt}^*$ equals one if a worker experiences a separation with $\text{WorkerSep}_{ijkt}^* \geq 0$ and zero otherwise. We consider three types of separations: (i) all firm-worker separation; (ii) an involuntary worker separation or firm layoff; and (iii) a voluntary worker separation or quit. The model specification includes individual, firm and industry specific control variables: (i) B_{it} is a set of worker characteristics, such as an age polynomial, interactions of a female dummy variable with age variables, marital status, tenure, region of residence dummy variables, earnings in period year $t - 1$, union membership dummy variable and interactions with the shutdown rates; (ii) C_{kt} are firm size dummy categorical variables; (iii) I_{jt} are a set of industry dummy variables; and (iv) D_t is a set of year dummy variables.

The analysis focuses on worker separations occurring at continuing firms. By definition, all workers at a firm experience a separation when their firm shuts down. We do not observe outcomes, separation and type of separation, for these workers if their firm continued operations. We account for the potential sample selection bias due to non-random firm shutdown using a bivariate probit model (BPWS) model, see Maddala (1983). The implication of this conditioning out of shutting down firms is that a finding of a positive relationship between industry shutdown rates and worker separation rates does not directly result from separations due to shutdown of individual firms. The selection equation describes the probability of a firm's shutdown and the outcome equation describes the probability of a worker separation:

$$\begin{aligned} \text{Firm Active}_{ijkt}^* &= \alpha^S + \beta^S X_{jt} + \gamma^S B_{it} + \sum_{k=1}^K \varphi_k^S C_{kt} + \sum_{j=1}^J \psi_j^S I_{jt} + \delta^S D_t \\ &\quad + \lambda RER_{jt} + v_{ijkt}, \end{aligned} \quad (10.2)$$

$$\begin{aligned} \text{Worker Sep}_{ijkt}^* &= \alpha^Q + \beta^Q X_{jt} + \gamma^Q B_{it} + \sum_{k=1}^K \varphi_k^Q C_{kt} + \sum_{j=1}^J \psi_j^Q I_{jt} \\ &\quad + \delta^Q D_t + u_{ijkt}. \end{aligned} \quad (10.3)$$

$$v_{ikjt}, u_{ikjt} \sim N(\mu, \Sigma), \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

Again, $\text{WorkerSep}_{ijkt}^*$, equals one if a worker experiences a separation with $\text{WorkerSep}_{ijkt}^* \geq 0$ and zero otherwise. A second indicator variable, FirmActive_{ikjt} , equals one if a firm remains active with $\text{FirmActive}_{ikjt}^* \geq 0$ and zero otherwise. The parameter ρ captures the correlation between the error terms in the two equations. In the remainder of the paper, *active* or *active* firms refer to firms that do not experience a shutdown in period t .

One source of identification occurs through the correlation parameter, ρ , and the nonlinearities of the BPWS model. Further, the shutdown equation also includes the industry-level US-Canada real exchange rate (RER_{jt}) as a regressor, while the worker separation equation does not strengthen identification through an exclusion restriction. Most of Canada's import/export activity is trade with the United States. The US-Canada real exchange rate affects the profitability of Canadian firms as the competitiveness of Canadian firms both in Canada and exporting to the US varies with the exchange rate between the two countries. Therefore, movements of the RER_{jt} likely affects a Canadian firm's shutdown or continue decision. Campa and Goldberg (2001) provide evidence that variations in RER have negligible effects on employment and number of jobs. This evidence provides support for the RER_{jt} serving as a valid exclusion restriction. The RER_{jt} variable is constructed according to the following formula $RER_{jt} = P_{jt}^{US}/P_{jt}^{CDN} \times e_t$, where P_{jt}^{US} is the US industry gross output price index, P_{jt}^{CDN} is the Canada industry gross output price index and e_t is the nominal bilateral exchange rate between Canadian and US in year t .

10.4 Results

This section considers the impact of industry shutdown rates on worker separation rates, layoff rates and quit rates in Tables 10.2, 10.3 and 10.4, respectively. In each table, the second column provides coefficient estimates and third column provides marginal effects for estimation results using only the subsample of workers at continuing firms without accounting for firm selection or Eq. (10.1). Columns four-six provide the results for the full selection model accounting for selection of continuing firms or Eqs. (10.2) and (10.3). Column four provides coefficient estimates for the firm probit selection equation, with column five providing coefficient estimates for the worker probit equation with associated marginal effect given in column six.

Table 10.2 Probability of a separation

Variable	Probit model		Bivariate probit with selection		
	Separations		Survival	Separations	
	Coef.	M.E.		Coef.	M.E.
Age	0.553***		-0.063*	0.548***	
Age ² /10	-0.201***		0.019	-0.199***	
Age ³ /100	0.031***		-0.002	0.031***	
Age ⁴ /1000	-0.002***		0.000	-0.002***	
Age total		0.001			0.001
Female*age	0.006**	0.002	-0.007	0.007***	0.002
Female*age ² /10	-0.003*	-0.001	0.003	-0.003*	-0.001
Female*age ³ /100	0.001	0.000	0.000	0.001	0.000
Female*age ⁴ /1000	0.000	0.000	0.000	0.000	0.000
Married	-0.093***	-0.033	0.025***	-0.093***	-0.034
Second job	0.326***	0.126	-0.311***	0.340***	0.132
Tenure	-0.034***	-0.013	0.014***	-0.035***	-0.013
Lagged earnings	-0.148***	-0.054	0.004***	-0.145***	-0.054
Atlantic	0.402***	0.156	0.090***	0.393***	0.153
Quebec	0.165***	0.062	0.052***	0.161***	0.062
Prairies	-0.005***	-0.002	0.070***	-0.009***	-0.003
BC	-0.008***	-0.003	0.059***	-0.011***	-0.004
Firm size < 5	-0.389***	-0.129	-0.997***	-0.276***	-0.097
Firm size 5–19	-0.181***	-0.064	-0.391***	-0.152***	-0.055
Firm size 20–49	-0.056***	-0.020	-0.149***	-0.047***	-0.017
Firm size 100–199	0.040***	0.015	0.072***	0.037***	0.014
Firm size 200–499	0.019***	0.007	0.136***	0.012***	0.005
Firm size > 500	-0.045***	-0.017	0.692***	-0.064***	-0.023
Union	0.425***	0.165	-0.176***	0.434***	0.170
Shutdown rate	1.560***	0.574	-4.537***	1.795***	0.669
Union*shutdown	-1.224***	-0.450	1.116***	-1.308***	-0.487
Real exchange rate			-0.093***		
$\rho(u^V, u^S)$			-0.455***		
Prob(y = 1)		0.344			
log likelihood	-3,901,850		-4,788,614		
Observations	6,938,850		7,186,450	Censored	247,600

Note: Coef. and M.E. denote the coefficients and marginal effects, respectively. Standard errors are reported in parentheses. *, **, and *** indicates statistical significance at the 10, 5, and 1 % levels, respectively. The first set of estimates are from a probit specification while the second set are from a bivariate probit with selection. The selection equation or Survival controls for the probability of firm survival while the outcome equation is the probability of a quit. The industry real exchange rate is the exclusion restriction

Table 10.3 Probability of a layoff or involuntary separation

Variable	Probit model		Bivariate probit with selection		
	Layoffs		Survival	Layoffs	
	Coef.	M.E.		Coef.	M.E.
Age	0.506***		-0.055	0.507***	
Age ² /10	-0.175***		0.016	-0.175***	
Age ³ /100	0.027***		-0.002	0.027***	
Age ⁴ /1000	-0.001***		0.000	-0.001***	
Age Total		0.003			0.003
Female*age	-0.053***	-0.017	-0.008*	-0.053***	-0.017
Female*age ² /10	0.028***	0.009	0.004	0.028***	0.009
Female*age ³ /100	-0.005***	-0.002	-0.001	-0.005***	-0.002
Female*age ⁴ /1000	0.000***	0.000	0.000	0.000***	0.000
Married	-0.079***	-0.024	0.030***	-0.079***	-0.024
Second job	0.022***	0.007	-0.283***	0.026***	0.008
Tenure	-0.029***	-0.009	0.015***	-0.029***	-0.009
Lagged earnings	-0.138***	-0.044	0.000	-0.138***	-0.044
Atlantic	0.660***	0.244	0.062***	0.659***	0.244
Quebec	0.216***	0.073	0.032***	0.216***	0.073
Prairies	-0.111***	-0.034	0.074***	-0.111***	-0.034
BC	0.013***	0.004	0.056***	0.013***	0.004
Firm size < 5	-0.051***	-0.016	-1.003***	-0.029***	-0.009
Firm size 5–19	0.013***	0.004	-0.391***	0.018***	0.006
Firm size 20–49	0.014***	0.005	-0.147***	0.016***	0.005
Firm size 100–199	-0.002	-0.001	0.072***	-0.002	-0.001
Firm size 200–499	-0.031***	-0.010	0.137***	-0.033***	-0.010
Firm size > 500	-0.176***	-0.052	0.687***	-0.180***	-0.053
Union	0.578***	0.211	-0.158***	0.580***	0.212
Shutdown rate	2.072***	0.655	-4.510***	2.130***	0.676
Union*shutdown	-1.150***	-0.363	0.984***	-1.168***	-0.371
Real exchange rate			-0.107***		
$\rho(u^V, u^S)$			-0.101***		
Prob(y = 1)		0.247			
log likelihood	-2,520,180		-3,407,726		
Observations	6,938,850		7,186,450	Censored	247,600

Note: Coef. and M.E. denote the coefficients and marginal effects, respectively. Standard errors are reported in parentheses. *, **, and *** indicates statistical significance at the 10, 5, and 1 % levels, respectively. The first set of estimates are from a probit specification while the second set are from a bivariate probit with selection. The selection equation or Survival controls for the probability of firm survival while the outcome equation is the probability of a quit. The industry real exchange rate is the exclusion restriction

Table 10.4 Probability of a quit or voluntary separation

Variable	Probit		Bivariate probit with selection		
	Quits		Survival	Quits	
	Coef.	M.E.		Coef.	M.E.
Age	0.086***		-0.054	0.088***	
Age ² /10	-0.030***		0.016	-0.031***	
Age ³ /100	0.004**		-0.002	0.004**	
Age ⁴ /1000	-0.000*		0	-0.000*	
Age total		-0.001			-0.001
Female*age	0.024***	0.002	-0.008*	0.024***	0.002
Female*age ² /10	-0.018***	-0.001	0.005	-0.018***	-0.001
Female*age ³ /100	0.004***	0.000	-0.001	0.004***	0.000
Female*age ⁴ /1000	-0.000***	0.000	0	-0.000***	0.000
Married	-0.047***	-0.003	0.029***	-0.048***	-0.003
Second job	0.569***	0.063	-0.287***	0.578***	0.067
Tenure	-0.050***	-0.003	0.015***	-0.050***	-0.003
Lagged earnings	-0.065***	-0.004	0	-0.065***	-0.004
Atlantic	-0.210***	-0.011	0.055***	-0.211***	-0.012
Quebec	-0.038***	-0.002	0.028***	-0.039***	-0.003
Prairies	0.146***	0.011	0.071***	0.143***	0.011
BC	-0.014***	-0.001	0.054***	-0.016***	-0.001
Firm size < 5	-0.565***	-0.022	-1.004***	-0.485***	-0.021
Firm size 5–19	-0.261***	-0.013	-0.392***	-0.239***	-0.013
Firm size 20–49	-0.082***	-0.005	-0.148***	-0.075***	-0.005
Firm size 100–199	0.037***	0.002	0.072***	0.034***	0.002
Firm size 200–499	0.025***	0.002	0.136***	0.020***	0.001
Firm size > 500	-0.021***	-0.001	0.685***	-0.036***	-0.002
Union	0.064***	0.004	-0.146***	0.071***	0.005
Shutdown rate	0.165*	0.011	-4.457***	0.341***	0.023
Union*shutdown	-1.084***	-0.071	0.923***	-1.124***	-0.077
Real exchange rate			-0.112***		
$\rho(u^V, u^S)$			-0.272***		
Prob(y = 1)		0.029			
log likelihood	-1,556,269		-2,443,628		
Observations	6,938,850		7,186,450	Censored	247,600

Note: Coef. and M.E. denote the coefficients and marginal effects, respectively. Standard errors are reported in parentheses. *, **, and *** indicates statistical significance at the 10, 5, and 1 % levels, respectively. The first set of estimates are from a probit specification while the second set are from a bivariate probit with selection. The selection equation or Survival controls for the probability of firm survival while the outcome equation is the probability of a quit. The industry real exchange rate is the exclusion restriction

10.4.1 Worker Separations

Table 10.2 presents estimation results considering all worker separations. The model estimates suggest that controlling for individual- and firm-specific characteristics, industry shutdown rates have a positive and statistically significant effect on the probability of a separation. The exclusion restriction variable RER_{jt} in the full selection model is highly significant and the correlation coefficient is negative but also highly significant. The effect of shutdown rates is negative as the probability of a firms shutdown is higher in industries with higher shutdown rates. There are changes in coefficient estimates when accounting for firm selection, especially the coefficient on industry separation rates. Looking at the marginal effects estimated for the BPWS model accounting for selection, a 1% rise in the industry shutdown rate causes a 0.66% rise in worker separation rates.⁴ However, unionization mutes the impact of industry shutdown as the marginal effect of the interaction variable (union \times shutdown rate) is negative at -0.487, which implies a 1% rise in industry shutdown rates leads to a 0.18% rise in worker separation rates at unionized firms. Higher firm instability leads to higher worker turnover in surviving firms but in unionized firms the effect of firm instability is considerably smaller.

The estimates on the other control variables are as follows. The probability of separation increases with age but at a declining rate. The total marginal effect with respect to age is (0.001).⁵ Individual characteristics have expected effects: marriage (-0.034), longer tenure (-0.013) and higher earnings in previous years (-0.054) can be expected to reduce the probability of a separation. Workers in the Atlantic Provinces (0.153) and Quebec (0.062) are more likely to experience separation than those in Ontario, whereas workers in the Prairies and British Columbia (-0.003) are less likely.

The estimates of the effects of firm-specific characteristics suggest that there is a U-shape relationship between the firm size and the probability of a separation. Those who work in mid-size firms (50–499 employees) are more likely to experience a separation than those who work in smaller firms and firms that employ more than 500 workers.

⁴The marginal effect on the industry shutdown rate is an elasticity since the probit models a probability and shutdown rates are probabilities as well. We report results in terms of marginal effects estimated at the means of the continuous explanatory variables and with dummy variables set to zero.

⁵ $\frac{\partial \phi(\hat{\theta}^S z^S)}{\partial age}|_{Age=\bar{Age}} = \phi(\hat{\theta}^S z_B^S)(\hat{\gamma}_1 A + 2\hat{\gamma}_2 \bar{Age} + 3\hat{\gamma}_3 \bar{Age}^2 + 4\hat{\gamma}_4 \bar{Age}^3)$, where z_B^S is a set of benchmark characteristics described in the previous footnote and ϕ is a standard normal density function.

10.4.2 Involuntary Separations or Firm Layoffs

The results for layoffs or involuntary worker separations are shown in Table 10.3. The parameter estimates in the selection equation of the BPWS model are very similar to those presented in Sect. 10.4.1, which is not surprising given that the probability of a firm continuing should not vary from across the models with different dependent variables in the outcome equations.

The effect of industry shutdown rates on the probability of a layoff is positive and highly significant in both the benchmark and BPWS models. Focusing on the BPWS model estimates, the estimated elasticity of worker layoff rates to industry shutdown rates is 0.68. Similar to the separation models, union membership dampens the effect of industry instability, so that a 1 % rise in industry shutdown rates leads to 0.31 % increase in layoff rates at unionized firms.

As with separations, the exclusion restriction variable RER_{jt} is highly significant, and the correlation coefficient is negative and highly significant as well. The effects of other variables are similar to the results for separations, which likely reflects the fact that over 70 % of worker separations are due to firm layoffs.⁶ All in all, these results show that most separations are involuntary, and the probability of a separation very much depends on the likelihood of being laid-off. The results also show that firm instability has a strong impact on layoffs in continuing firms. In particular, higher shutdown rates in an industry increase the probability of layoffs in surviving firms in that industry.

10.4.3 Voluntary Separations or Worker Quits

The results in Table 10.4 show that controlling for individual and firm-specific characteristics, industry shutdown rates have a positive and significant effect on the probability of a quit; the estimated marginal effect or elasticity is 0.02 in BPWS model. However, the results reverse qualitatively for unionized firms. A 1 % increase industry shutdown rates reduces the probability of quits in unionized firms by 0.054 %. The estimated effects of individual characteristics (age, sex, marital status, tenure and lagged earnings) are in line with other studies on job separations. Individuals are substantially more likely to quit secondary employment (0.085 for BPWS). With respect to firm characteristics, the probability of a quit increases with firm size for small and mid-size firms (<200 employees). For larger firms, the opposite is true. Workers in Atlantic Provinces are least likely to quit (-0.012), whereas workers in the Prairies are most likely (0.011). The probability of a quit is lower for those with longer tenure (-0.003) and higher earnings in previous years (-0.004) but higher in the case of a second job (0.067).

⁶The baseline probability of a worker separation is 34.4 %, while the baseline probability of a firm layoff is 24.7 %.

10.5 Conclusions

Our findings underscore the complexity of the issue of individual job stability. Much of the attention in the literature has been paid to the causes and consequences of worker separations. This paper focuses on the impact of industry instability on worker separations. As such, we highlight a less obvious but also important relationship between industry instability and quits. The results of the analysis do not invalidate the hypothesis that higher industry shutdown rates lead to greater worker turnover. Industry shutdown rates have a positive impact on overall worker separation rates, involuntary layoff rates and voluntary quit rates in firms that remain active. Thus, workers are more likely to be laid off but also are more likely to quit in anticipation of future layoffs. Such separations are “voluntary” in a narrower sense than is usually assumed, and their long-run effects on individual earnings may be similar to the effects of layoffs. Our results suggest that worker separations move in the same direction for firms within an industry.⁷ Deteriorating industry conditions require firms to reduce costs with one possibility being worker shedding, either voluntary or involuntary. At the extreme, worker shedding leads to firm shutdown. Such possibility is the subject of our future research.

Finally, we highlight the fact that industry instability has a much bigger impact on involuntary separation rates than voluntary separation rates. We find statistically significant effects, but the economic significance of industry shutdown rates on quit rates is small. Thus, industry instability increases worker separations mainly through worker layoffs rather than worker quits. Lise et al. (2013) demonstrate the process of on the job search by currently employed workers. Although examining voluntary worker separations remains a relevant question, these results suggest that involuntary separations are more relevant to job turnover and worker outcomes. Huynh et al. (2014) provide such an avenue. Their analysis suggests that higher industry instability affects workers both from an extensive margin through higher permanent layoffs and an intensive market through lower wages. This paper provides a path for future research to examine worker outcomes of quits and layoffs in relation to worker wage performance.

Acknowledgements Huynh and Voia gratefully acknowledges the assistance and hospitality of Statistics Canada Economic and Social Analysis Divisions. The authors “Kim P. Huynh, Yuri Ostrovsky, Robert J. Petrunia, and Marcel C. Voia” are especially indebted to Leonard Landry for his gracious efforts on data management. We thank Michael Veall, Iourii Manovskii, Gueorgui Kambourov, John Stevens, James Townsend and participants of the 2009 Cornell VRDC conference and 2010 Canadian Economics Association for comments and suggestions. The views expressed in them are those of their authors and not necessarily the views of the Bank of Canada or Statistics Canada. All errors and opinions are our own.

⁷We thank Michael Veall for this observation.

References

- Beach CM, Finnie R, Gray D (2003) Earnings variability and earnings instability of women and men in Canada: how do the 1990s compare to the 1980s? *Can Public Policy* 29(s1):41–64
- Browning M, Lusardi A (1996) Household saving: micro theories and micro facts. *J Econ Lit* 34(4):1797–1855
- Campa JM, Goldberg LS (2001) Employment versus wage adjustment and the U.S. dollar. *Rev Econ Stat* 83(3):477–489
- Fraser CD (2001) Income risk, the tax-benefit system and the demand for children. *Economica* 68(269):105–125
- Gottschalk P, Moffitt R (1994) The growth of earnings instability in the U.S. labor market. *Brook Pap Econ Act* 25(2):217–272
- Gottschalk P, Moffitt R (2009) The rising instability of U.S. earnings. *J Econ Perspect* 23(4):3–24
- Guiso L, Jappelli T, Pistaferri L (2002) An empirical analysis of earnings and employment risk. *J Bus Econ Stat* 20(2):241–53
- Huynh KP, Jacho-Chávez DT, Petrunia RJ, Voia M (2011) Functional principal component analysis of density families with categorical and continuous data on Canadian entrant manufacturing firms. *J Am Stat Assoc* 106(495):858–878
- Huynh KP, Ostrovsky Y, Petrunia RJ, Voia MC (2015) Industry shutdown rates and permanent layoffs: evidence from firm-worker matched data. mimeo
- Huynh KP, Petrunia RJ (2010) Age effects, leverage, and firm growth. *J Econ Dyn Control* 34(5):1003–1013
- Huynh KP, Petrunia RJ, Voia M (2010) The impact of initial financial state on firm duration across entry cohorts. *J Ind Econ* 58(3):661–689
- Jacobson LS, LaLonde RJ, Sullivan DG (1993) Earnings losses of displaced workers. *Am Econ Rev* 83(4):685–709
- Kambourov G, Manovskii I (2009) Occupational specificity of Human capital. *Int Econ Rev* 50(1):63–115
- Lise J, Meghir C, Robin J-M (2013) Mismatch, sorting and wage dynamics. NBER Working papers 18719, National Bureau of Economic Research, Inc.
- Maddala G (1983) Limited dependent and qualitative variables in econometrics. Cambridge University Press, Cambridge
- Morissette R, Ostrovsky Y (2005) The instability of family earnings and family income in Canada, 1986–1991 and 199–2001. *Can Public Policy* 31(3):273–302
- Pistaferri L (2003) Anticipated and unanticipated wage changes, wage risk, and intertemporal labor supply. *J Labor Econ* 21(3):729–728
- Quintin E, Stevens JJ (2005) Raising the bar for models of turnover. Finance and economics discussion series 2005–23, Board of Governors of the Federal Reserve System (U.S.)

Chapter 11

Inputs, Productivity and Agricultural Growth in Sub-Saharan Africa

Alejandro Nin-Pratt

Abstract This study employs a growth accounting approach to revisit past performance of agriculture in sub-Saharan Africa (SSA) and to analyze the relationship between the input mix used by SSA countries and productivity levels observed in the region. Findings show that improved technical efficiency has been the main driver of growth in recent years benefiting poorer, low labor productivity countries. Countries with higher output and input per worker have benefited much more from technological progress than poorer countries, suggesting that technical change has done little to reduce the gap in labor productivity between countries. Results also show that the levels of input per worker used in SSA agriculture at present are extremely low and associated with less productive technologies, and that technical change has shifted the world technological frontier unevenly, increasing the distance between SSA countries and those countries with the “right” input mix.

Keywords Agriculture • Appropriate technology • Total factor productivity • Africa south of the Sahara

JEL code: O13, O33, O55, Q16, Q18

11.1 Introduction

The evidence of improved performance of agriculture in Sub-Saharan Africa in recent years has indeed been quite striking when compared with the past. For the first time, the sector has maintained a real growth rate of 3.4 % per year, well above a population growth rate of 2.5 %. Recent studies (e.g. Alene 2010; Block 1995, 2010; Fuglie 2011; Fuglie and Rada 2012; Nin-Pratt and Yu 2012) have shown how

A. Nin-Pratt (✉)

Environment and Production Technology Division, International Food Policy Research Institute, 2033 K Street NW, Washington, DC 20006, USA

e-mail: a.ninpratt@cgiar.org

African agricultural performance improved in the aftermath of political, policy, and institutional reforms since the 1990s, increasing public investment and reducing the heavy taxation on agriculture.

Most of these studies concluded that the region should increase its efforts to accelerate total factor productivity (TFP) growth and technical change. For example, Fuglie and Rada (2012) showed that despite recent improvement, agricultural productivity growth in SSA continues to lag behind every other region of the world, growing at rates that are roughly half of the average rate of developing countries. They concluded that SSA should increase its accumulated knowledge capital from long-term national and international investments in agricultural R&D, which are gradually delivering improved technologies to farmers. At the same time, they highlight the need of strengthening the broader enabling environment for farmers to access technology, markets and the necessary support services for raising agricultural productivity in SSA. Similarly, Nin-Pratt and Yu (2012) concluded that several warning signs still exist, calling for more efforts to sustain TFP growth, arguing that without increases in the rate of growth of technical change, TFP growth is expected to slow down in the coming years as countries catch up with efficiency levels at the production frontier.

Acknowledging the need of accelerating technical change through increasing investments in agricultural R&D and improving the enabling environment for technology adoption, this study introduces a second dimension to the puzzle of SSA's agricultural growth: the role of the input mix and the need to increase capital and inputs per worker not only to boost output per worker but also to accelerate technology adoption and TFP growth. There are significant implications in terms of policy, allocation of R&D investment, the type of technologies to promote and the growth path that countries could follow depending on how we interpret the role of different factors of production and inputs in the process of technical change and their effect on productivity levels.

The level and combination of labor, capital and materials (e.g. fertilizer, feed) has not been central to the discussion of technical change and TFP growth in part because the conceptual framework normally used to analyze these issues assumes a uniform technology frontier for all countries. This means that a poor country using mostly labor and land and very little capital can produce similar levels of output per unit of aggregate input than richer countries using a different combination of inputs (e.g. high levels of capital per worker). If observed TFP levels in poor countries are low, these differences in productivity reflect inefficiencies that can be reduced if poor countries have access to technologies used by frontier countries.

Why do improved technologies not diffuse across borders allowing SSA countries to catch-up? The reason, according to this view, is that there may be barriers to the adoption of technology like those resulting from agroecological differences, institutional differences or inefficient social arrangements (e.g. lack of competitive markets) that result in lack of technologies due to barriers to adoption and the inefficient use of technologies already in place. If countries are able to reduce these barriers (e.g. by adapting technologies to their agroecologies and improving institutions and infrastructure) then TFP levels should converge to those of richer

countries. The assumption is that technical change is “neutral” so innovations from rich countries should benefit poor countries after some investment is done to adapt these technologies to a different economic environment, and that there are large gains to be made in terms of TFP by closing the technological gap even by poor countries at low levels of capitalization (Jerzmanowski 2007).

What happens if the technological frontier is not uniform and not every country can reach the same TFP level? Or as Jerzmanowski (2007) puts it, what if countries choose the best technologies available to them but their choice is limited by the fact that not all existing technologies are equally suited to every economy? In this case the relevant question is: what determines whether certain technologies are appropriate for a particular economy?

One possible answer to this question is provided by the literature on appropriate technology which argues that depending on the country’s relative stocks of physical and human capital, some technologies may be more or less productive than others. Formally, this means that TFP is a function of factor endowments. Recent theoretical contributions have emphasized the potential dependence of productivity on inputs such as physical or human capital, invoking the appropriate technology paradigm to explain differences in income levels and the lack of convergence (e.g. Basu and Weil 1998; Acemoglu and Zilibotti 2001). In these models, rich countries invent technologies that are compatible with their own factor mix, but these technologies do not work well with the very different factor mix of poor countries. For example, in the case of agriculture one can think of technologies requiring the intensive use of capital (irrigation and mechanization), or an appropriate match of land and machines (e.g. tractors). Consequently these technologies are “inappropriate” for poor countries with high capital costs and when adapted to their economic conditions do not result in the same gains in TFP than those observed in richer countries.

In summary, the appropriate technology theory argues that low TFP is a result of the technology frontier being lower for some factor endowments. On the other hand, what Jerzmanowski (2007) calls the “efficiency” view maintains that the frontier is the same everywhere, but some countries operate below it. A better understanding of the role of inputs on TFP gaps could have important policy implications. Should SSA countries promote commercial agriculture so a group of producers could converge faster to production conditions in richer countries to overcome technology “inappropriateness”? Should governments invest in agricultural R&D to develop technologies appropriate for poor households producing with low levels of capital assuming that there are always new productive techniques to be developed? Or, is there a limit to increase productivity at a certain level of human and physical capital per worker especially if this level is very low, meaning that the oxcart can only be improved so much? Can we still expect large gains in productivity from improvements in efficiency and adoption of existing technologies? Is the slow pace of technology adoption and TFP growth in SSA the result of inappropriateness of technology given the very particular conditions and low levels of capitalization of agriculture in these countries? Looking for answers to these questions is important because different strategies can have different costs in terms of investment, time and welfare for SSA economies.

Unlike previous papers looking at agricultural growth in SSA, in this study we decompose levels and growth in output per worker using a growth accounting approach to analyze the explanatory power of the efficiency versus the appropriate technology hypothesis to explain productivity differences between SSA and other regions. Efficiency measures together with econometric estimates of input elasticities of a Cobb-Douglas production function are the main components of our model. In the next section we present the conceptual framework and approach followed in this study. Section 11.3 describes the data used, technical aspects and results of the efficiency and input elasticities estimation. Section 11.4 revisits the analysis of past performance of agriculture in SSA looking at growth of output per worker and its decomposition into efficiency, technical change and input growth. Section 11.5 shows results of the decomposition of the levels of output per worker into levels of efficiency, technology and inputs and how the efficiency and appropriate technology explain differences in labor productivity and TFP levels. The last section concludes and derives policy implications.

11.2 Conceptual Framework and Approach

An accepted view on the analysis of agricultural productivity adopts a Cobb-Douglas production function with constant returns to scale to estimate TFP assuming that countries have access to a common technology represented as $y = A\pi x^\alpha$ where y and x are respectively output and input per worker and A represents TFP or the part of output not explained by inputs x . This view implies a uniform technology frontier for all countries, that is, all countries face the same “ A ” in the production function and differences in TFP reflect inefficiency or a gap from the frontier due to barriers to the adoption of technology, natural resources, lack of competitive markets or other efficient social arrangements (Jerzmanowski 2007).

An alternative view to the standard growth accounting analysis asserts that the technology frontier is not uniform (i.e., not every country faces the same A in the production function above) and that countries choose the best technologies available to them, however, their choice is limited by the fact that not all existing technologies are equally suited to every economy. One explanation for this is that appropriateness depends on the mix of inputs. That is, depending on the country’s relative stocks of labor, skills and physical capital, some technologies may be more or less productive than others. Under this assumptions the A in the Cobb-Douglas production function becomes $A = A(x)$ (Jerzmanowski 2007). As discussed in Basu and Weil (1998) and Acemoglu and Zilibotti (2001), the appropriate technology paradigm explain differences in income levels and the lack of convergence. For example, in the paper by Acemoglu and Zilibotti (2001), rich countries invent technologies that are compatible with their own factor mix, but these technologies do not work well with the very different factor mix of poor countries and consequently the most productive technologies are inappropriate for developing countries and, even if adopted, do not raise their TFP levels.

In this section we present a model of appropriate technology adapted from Jerzmanowski (2007), which is part of a large literature that examines barriers to the transfer of technology across countries including Basu and Weil (1998); Parente and Prescott (1994); Segerstrom et al. (1990); Grossman and Helpman (1991) and Barro and Sala-i-Martin (1991) among others. We start by presenting the basic elements of the growth accounting method followed by the nonparametric approach to productivity analysis. We then combine elements of these two approaches to define a “hybrid” model where the Cobb-Douglas production function is defined as a frontier function and TFP is decomposed into an efficiency component which is independent of the level of inputs and a technology component expressed as a function of input per worker.

11.2.1 Growth Accounting Approach

Much of the literature on agricultural productivity assumes a Cobb-Douglas production function with constant returns to scale to estimate TFP since the seminal agricultural studies by Griliches (1964) and Hayami and Ruttan (1985). Eberhardt and Teal (2013) review of the literature refers to several studies applied to agriculture using the Cobb-Douglas function including Craig et al. (1997); Cermeno et al. (2003); Bravo-Ortega and Lederman (2004) and Fulginiti et al. (2004). Recent work looking specifically to SSA agriculture includes Block (2010) and Fuglie (2011). Under this approach, the output per worker in country i is given by

$$y_i = F_i(x) = A_i \prod x_{ij}^{\alpha_j} \quad (11.1)$$

where y_i is agricultural output per worker, x_{ij} is a set of j observed inputs per worker and A_i is unobserved TFP with technology parameters α_j constant over time. The production function shifter A_i can be modeled borrowing from Fuglie (2011) as:

$$\ln(A_i) = T_i + \eta_i + \sum \beta_{ki} Z_{ki} + \varepsilon_i \quad (11.2)$$

where T represents technology levels, η_i is a random and unobserved country specific effect and Z_{ki} are observed differences in resource quality while ε_i is a random component capturing measurement error. Changes in A_i over time shift the production function and are interpreted as factor-neutral improvements in technology or production efficiency.

As discussed in Fuglie (2011), production elasticities α_j can be interpreted as the share of output that each input receives in payment for its contribution to the production process and under certain assumptions this shares indicate the payments that the owners of these resources receive when inputs are paid their value-marginal product. In this way, econometric estimation of the parameters of the production

function are used instead of input prices, which are normally not available, to define TFP and an index of TFP growth expressed in terms of growth rates:

$$\ln(TFP_i) = \ln(y_i) - \sum \alpha_j \ln(x_{ij}) \quad (11.3)$$

$$T\dot{F}P = \dot{Y}_i - \sum \alpha_j \dot{x}_{ij} \quad (11.4)$$

One of the disadvantages of this approach is that it involves strong technical and economic assumptions, like profit maximization and imposing a functional form. On the other hand, Fuglie (2011) argues that imposing more structure could be an advantage when dealing with data with a high degree of measurement error as it can help produce more plausible results.

11.2.2 Non-Parametric Approach

The nonparametric approach known as Data Envelopment Analysis (DEA) has become especially popular because it is easy to compute and does not require information about input or output prices or assumptions regarding economic behavior, such as cost minimization and revenue maximization. The method has been extensively applied to the international comparison of agricultural productivity. See, for example, Bureau et al. (1995), Fulginiti and Perrin (1997), Lusigi and Thirtle (1997), Prasada Rao and Coelli (1998), Arnade (1998), Fulginiti and Perrin (1999), Chavas (2001), Suhariyanto et al. (2001), Suhariyanto and Thirtle (2001), Trueblood and Coggins (2003), Nin et al. (2003), Ludena et al. (2007), Alene (2010), and Nin-Pratt and Yu (2012).

In general, the nonparametric approach assumes that agricultural output per worker in country i is given by a production function of the form:

$$y_i = E_i \times F(x) \quad (11.5)$$

where y is output per worker, x is a vector of inputs used in production and E measures efficiency in the use of inputs and takes values between zero and one. The production function $F(x)$ satisfies free disposal and constant returns to scale and represents the production possibility frontier or the maximum attainable output given inputs. Actual output y results from the product of potential output and efficiency. In this context the production set S is defined as:

$$S = \{(x, y) : y \leq F(x)\} \quad (11.6)$$

The output distance function $D(x, y)$ expresses the maximum proportional expansion of output given inputs, or the maximum increase in output (within S) given that inputs remain constant, which is captured by θ as follows:

$$D(x, y) = [\sup \{\theta : (x, \theta y) \in S\}]^{-1} \quad (11.7)$$

where $D(x, y) \leq 1$ if and only if $(x, y) \in S$, and $D(x, y) = 1$ implies that production takes place on the technological frontier. The distance function for a particular country i^* is estimated using linear programming as described in Sect. 11.3.

Growth in output per worker between periods 0 and 1 can be represented, adapting notation from Kumar and Russell (2002) as:

$$\frac{y_1}{y_0} = \frac{E_1 \times F_1(x_1)}{E_0 \times F_0(x_0)} \quad (11.8)$$

where y_1 and y_0 represent output per worker in the final and initial period respectively, $F_1(x_1)$ is potential output that can be achieved using technology of the final period and the amount of inputs used in that same period and E_1 is efficiency of country i in the final period. Multiplying top and bottom by $F_0(x_1)$ or potential output that can be obtained using the technology of the initial period with inputs used in the final period we obtain:

$$\frac{y_1}{y_0} = \frac{E_1}{E_0} \times \frac{F_1(x_1)}{F_0(x_1)} \times \frac{F_0(x_1)}{F_0(x_0)} \quad (11.9)$$

Equation (11.9) is a decomposition of change in labor productivity between two periods for country i . The first term in the right hand side is the change in efficiency or the change in the distance to the frontier; the second term is the shift of the frontier between the two periods measured relative to the coordinates of country i in output space in the final period (potential output is measured with respect to x_1); and the last term is a measure of the change in potential output as a result of a change in the level of inputs, or movement along the frontier in the initial period.

The effect of changes in technology and inputs is path dependent which means that we can build a similar index by multiplying top and bottom in (11.8) by $F_1(x_0)$ instead of using $F_0(x_1)$ as before to obtain:

$$\frac{y_1}{y_0} = \frac{E_1}{E_0} \times \frac{F_1(x_0)}{F_0(x_0)} \times \frac{F_1(x_1)}{F_1(x_0)} \quad (11.10)$$

In this case, the shift in the frontier is measured with respect to country i 's coordinates in the production space in the initial period and the last term represents movement along the frontier in the final period. Expressions (11.9) and (11.10) are equal only when technological change is Hicks neutral in which case the shift in the frontier is independent of the value of the input-labor ratio. To avoid the problem of path dependence, Caves et al. (1982) adopted the “Fisher ideal” decomposition based on the geometric averages of the two measures of the effects of technological change and capital accumulation multiplying top and bottom of (11.9) by $[F_1(x_0) F_0(x_1)]^{1/2}$:

$$\frac{y_1}{y_0} = \frac{E_1}{E_0} \times \left[\frac{F_1(x_1)}{F_0(x_1)} \times \frac{F_1(x_0)}{F_0(x_0)} \right]^{1/2} \times \left[\frac{F_0(x_1)}{F_0(x_0)} \times \frac{F_1(x_1)}{F_1(x_0)} \right]^{1/2} \quad (11.11)$$

This approach has the advantage of imposing minimum restrictions on the production structure. On the other hand, because of its deterministic character it is not possible to evaluate the precision of the predicted efficiency levels if inputs and outputs are subject to stochastic variation. As the method constructs the production frontier based on efficient points it is naturally sensitive to outliers and measurement error.

11.2.3 The “Hybrid” Approach: Appropriate Technology

This approach goes along the lines of neoclassical growth accounting in defining TFP growth as the ratio of output and input growth, with the aggregate production function being defined as Cobb-Douglas with CRS. Within this neo-classical framework, it also disentangles technical change along the technological frontier from changes in technical efficiency. Starting from (11.5), we impose the Cobb-Douglas functional form to the generic expression $F(x)$ representing potential output:

$$y_i = E_i \times \left[T_i \prod x_{ij}^{\alpha_j} \right] \quad \text{where } A_i = E_i \times T_i \quad (11.12)$$

Notice that this hybrid model, unlike neoclassical growth accounting, deals exclusively with the best practice technology, not the average practice technology. In other words, the Cobb-Douglas production function is a frontier production function where TFP is decomposed into efficiency and available technology levels. Using growth accounting approach (dropping the country index) we can express the output growth decomposition between period 0 and 1 as:

$$\frac{y_1}{y_0} = \frac{E_1}{E_0} \times \frac{\bar{T}_1}{T_0} \times \prod \left(\frac{x_{j1}}{x_{j0}} \right)^{\alpha_j} \quad (11.13)$$

The expression in (11.13) is known in the growth accounting literature as the “appropriate technology vs. efficiency” output growth decomposition (Basu and Weil 1998; Jerzmanowski 2007; Growiec 2012). This specification allows for two determinants of TFP differences: country-specific levels of efficiency and country-specific levels of available technology which is allowed to be factor specific: $T_i(x)$ (Fig. 11.1).

The left panel of Fig. 11.1 characterizes a model of production where all countries have access to the same technology represented by the production function $y = Ax^\alpha$. In this setting differences in output per worker between an efficient country (C2) and an inefficient country (C1) are explained by (a) TFP levels which result from inefficiency (measured as the distance of C1 to the frontier given the level of input x_1 used); and (b) by differences in the level of input x used (increasing inputs from x_1 to x_2 will reduce the difference in output per worker to differences in efficiency only).

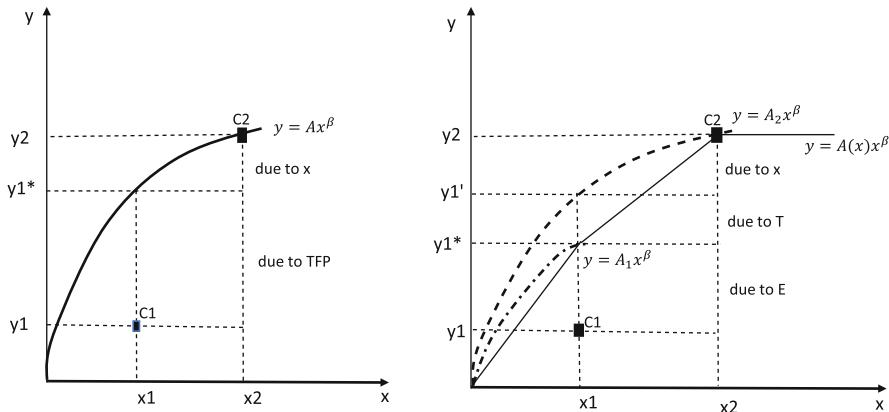


Fig. 11.1 Standard and appropriate technology levels accounting decomposition. Note: Left panel assumes that technology $y = Ax^\alpha$ is available to all countries and differences are due to input-labor level and TFP. Right panel: technology is a function of input per worker so country 1 cannot access country 2's technology. Source: Adapted from Jerzmanowski (2007)

The right panel in Fig. 11.1 represents production with appropriate technology. In this case, the true frontier is a function of input per worker. For each input-labor combination there is a particular production function (A is a function of x). The difference with the right panel is that in the left panel there is an intermediate level of output y_1' that C_1 cannot achieve with its present level of inputs. The difference $y_1' - y_1^*$ is due to appropriate technology. This means that to achieve productivity levels of C_2 , C_1 can increase efficiency up to certain point but to catch-up with C_2 , C_1 needs to increase input per worker to operate on C_2 production function and face TFP levels A_2 instead of A_1 .

The empirical application of the appropriate technology model used in this study implies the estimation of the global production frontier for agriculture using a DEA approach and the parameters of the Cobb-Douglas function, discussed in the next section.

11.3 Empirical Model and Implementation

Output and input data were collected from FAO (2014) covering a period of 50 years from 1961 to 2011. The final database includes 121 countries including 38 SSA countries, one output (total agricultural production), and six inputs (animal feed, fertilizer, labor, agricultural land, and crop and livestock capital). Output is defined as the value of gross agricultural production expressed in constant 2004–2006 International dollars (I\$). It includes crop and livestock production. Animal feed is the amount of edible commodities (cereals, bran, oilseeds, oilcakes, fruits, vegetables, roots and tubers, pulses, molasses, animal fat, fish, meat meal,

whey, milk, and other animal products from FAOSTAT food balance sheets) fed to livestock during the reference period. Quantities of the different types of feed are transformed into metric tons of maize equivalents using information of energy content for each commodity. Fertilizer is measured as the quantity of nitrogen, phosphorus, and potassium (N, P, K) in metric tons of plant nutrient consumed in agriculture by country and year as used in Fuglie and Rada (2012) available from the US Department of Agriculture (USDA 2014). Labor is total economically active population in agriculture engaged in or seeking work in agriculture, hunting, fishing, or forestry, whether as employers, own account workers, salaried employees, or unpaid workers. The data on labor are originally from FAO, which currently reports the number of economically active adults in agriculture from 1980 onward. For our analysis we used the labor data from USDA (2014) that uses annual growth rates from 1961 to 1979 previously reported by FAO to derive estimates for 1961–1979, extrapolating backward from FAO's 1980 figures. Labor figures for Nigeria were adjusted following Fuglie and Rada (2012) assuming 2 % annual growth in agricultural labor for subsequent years. Land includes land under temporary crops (doubled-cropped areas are counted only once); temporary meadows for mowing or pasture; land under market and kitchen gardens; land temporarily fallow (less than 5 years); and land cultivated with permanent crops such as flowering shrubs (coffee), fruit trees, nut trees, and vines; but excludes land under trees grown for wood or timber. Pasture land includes land used permanently (5 years or more) for herbaceous forage crops, either cultivated or growing wild (wild prairie or grazing land), all measured in hectares. As a measure of capital stock we use FAO's new series of capital stock covering the period 1975–2007 valued at 2005 constant prices as the base year, which was developed by multiplying unit prices by the quantity of physical assets “in use” compiled from individual countries. Capital used in crop production (crop capita) includes land development, irrigation works, structures and machinery. Livestock capital includes animal stock, structures for livestock and milking machines. As the capital series are available until 2007 we project them to 2011 using values of their different components from FAO: machinery, area of permanent crops, and animal stock.

Countries in our sample were grouped by agroecological zone (AEZ). These zones were defined based on data from Lee et al. (2005) and the table in the Appendix groups SSA countries by AEZ. Results and discussion in Sects. 11.4 and 11.5 focus on the period that goes from 1971 to 2011 covering the post-independence policies, the period of structural adjustment and the accelerated growth of recent years.

11.3.1 Efficiency Estimates

As discussed in Sect. 11.2, we use distance functions to measure output oriented technical efficiency for our sample of countries including information on agroecologies for the different countries to account in part for resource quality. We do

this in two steps. We first calculate distance functions pooling all countries in our sample to measure the distance of each country to the world frontier in each year. We then group countries by agroecology and estimate the distance of all countries to the frontier of their respective group. The distance function of a country in the k th group is defined as:

$$D^k(x, y) = [\sup \{\theta : (x, \theta y) \in S^k\}]^{-1} \quad (11.14)$$

Technical efficiency with respect to the world metafrontier is:

$$D^*(x, y) = [\sup \{\theta^* : (x, \theta^* y) \in S^*\}]^{-1} \quad (11.15)$$

The metafrontier envelopes the group frontiers which means that $D^k(x, y) \geq D^*(x, y)$ for all k . Following Rambaldi et al. (2007), we define the Technology Gap Ratio (TGR) in year t as the ratio of the two distances:

$$TGR^k = \frac{D(x, y)^*}{D(x, y)^k} \leq 1 \quad (11.16)$$

Rearranging terms, we define the distance to the metafrontier as the product of the technology gap between group k 's frontier and the metafrontier (TGR^k) and distance to the group's frontier:

$$D^*(x, y) = TGR^k \times D^k(x, y) \quad (11.17)$$

To estimate the distance function for a particular country io with respect to the world metafrontier we solve the following linear programming problem:

$$\begin{aligned} D^*(x_{io}, y_{io}) &= \max_{\theta^*, \lambda} \theta^*_{io} \\ \text{Subject to : } \theta^*_{io} y_{io} &\leq \sum_{i=1}^I \lambda_i y_i \text{ and} \\ x_{io,j} &\geq \sum_{i=1}^I \lambda_i x_{i,j} \text{ for inputs } j = \{1, \dots, J\}, \\ \lambda_i &\geq 0 \end{aligned} \quad (11.18)$$

On average, technical efficiency of SSA countries is low but improved in the last decade. The technology gap of the region is about 75 % which means that TFP level at the frontier of the agroecologies where SSA countries produce is 25 % lower than TFP level at the metafrontier. This distance has reduced in recent years.

11.3.2 Input Elasticities and the Cobb-Douglas Production Function

The empirical framework to estimate input elasticities of the Cobb-Douglas production function follows Eberhardt and Teal (2013) and builds on a common factor representation of the log-linearized production function, allowing to accommodate nonstationarity and correlation across panel members. Borrowing notation from Eberhardt and Teal (2013) we represent the Cobb-Douglas production function in logs as:

$$y_{it} = \beta'_i x_{it} + \mu_{it} \quad (11.19)$$

$$\mu_{it} = \alpha_i + \lambda'_i f_t + \varepsilon_{it} \quad (11.20)$$

The Cobb-Douglas production function (11.19) has observed output (y_{it}) and observed inputs (x_{it}) including labor, crop capital stock, livestock capital stock, fertilizer, feed and agricultural land (all in logarithms). The constant term is represented by a combination of country-specific effects (α_i) and a set of common factors f_t which can have different effects across countries (i).

The model allows for endogeneity as the input variables x_{it} are driven by a set of common factors g_{jt} and by the set (or subset) of factors f_t influencing output in (11.19) and (11.20), which means that some unobserved factors driving agricultural production are likely to drive, at least in part, the evolution of the inputs:

$$x_{ijt} = \pi_{ij} + \delta'_{ij} g_{jt} + \phi_{ij} f_t + v_{ijt} \quad (11.21)$$

Finally, (11.22) indicates that the latent factors are persistent over time, which allows for the setup to accommodate nonstationarity in factors ($\varrho = 1$, $\kappa = 1$).

$$f_t = \varrho' f_{t-1} + \epsilon_t \text{ and } g_t = \kappa' g_{t-1} + \epsilon_t \quad (11.22)$$

The parameter of interest for this study is the mean effect β . As in Eberhardt and Teal (2013) we consider different models to estimate β . These models deal with unobserved heterogeneity, cross section dependence and dependence due to latent common factors. We divide these models into two groups. Pooled models assume parameter homogeneity: all countries share the same slope parameters ($y_{it} = \beta' x$).

Within this group we estimated the Pooled Ordinary Least Squares model (POLS) with year dummy variables; the Two-way fixed effects model (2FE), including country and year dummy variables to capture country and year specific effects; the first-difference OLS model (FD-OLS), used to address the problem of omitted variables is obtained by running a pooled OLS estimation of the regression of the difference y : $y_{(t)} - y_{(t-1)}$ against y : $x_{(t)} - x_{(t-1)}$ wiping out time invariant

omitted variables; and the Pesaran (2006) common correlated effects (CCE) pooled estimator that uses the cross-section averages of the observed output and input variables (averages of y and x) as proxies for the latent factors f_t , assuming that unobserved factors which influence productivity are common to all countries. This model is extended as in Eberhardt and Teal (2013) using different weight-matrices to calculate the cross-section averages used as proxies for the latent factors f . That is, instead of assuming the same cross-section simple average to capture the impact of unobserved effects, different weights are used to calculate the average assuming that not all unobserved effects affect a particular country in the same way. The different versions of the CCE model are the following: CCEP-neighbor (CCEPn) using averages of contiguous neighbors for each country, assuming that common shocks between countries are transmitted only between neighboring countries; CCEP-distance (CCEPd) where cross-section averages are calculated using the inverse of the population weighted geographic distance between countries; and CCEP-cultivated land (CCEPc) where weights for every country pair are constructed based on the share of cultivated land within each of twelve climatic zones as defined in Jaffe (1986) and used in Eberhardt and Teal (2013), a more detailed climatic classification than the four agroecological zones defined here to control for natural resource quality in the efficiency comparisons.

The second group of models allows for heterogeneous slopes ($y_{it} = \beta_i'x$). These models are able to accommodate the type of endogeneity presented in the original model (11.19)–(11.22) to arrive at consistent estimates for common slope coefficients calculated as the mean of heterogeneous β_i . Simulations studies (for example, Coakley et al. 2006) show that results from these models are robust even when the cross-section dimension is small, when variables are non-stationary, and in the presence of weak unobserved common factors (spatial spillovers). Within this group we estimated the following models: Pesaran and Smith (1995) Mean Group (MG) where the intercepts, slope coefficients, and error variances are all allowed to differ across groups. The model assumes away cross-section dependence ($\lambda_i = 0$) and estimates separately individual country regressions. The heterogeneous version of the CCE models (CMG) estimates individual country regressions augmented by cross-section averages of dependent and independent variables using the data for the entire panel. As in the case of the CCE models, the neighbor, distance and cultivated land versions of the CMG model are defined by using different weights to calculate the cross-section averages. Finally, the augmented Mean Group estimator (AMG) of Eberhardt and Bond (2009), conceptually similar to the heterogeneous Mean Group version of Pesaran (2006) CCE estimator (CMG), is implemented in three steps: (a) a pooled regression model augmented with year dummies is estimated by first difference OLS and the coefficient on the year dummies are collected representing the common dynamic process affecting all countries; (b) the country specific regression model is then augmented with estimates from (a); finally in (c) country-specific parameters are averaged across the panel. A second version of this model (AMG2) imposes a unit coefficient in the dynamic process variable in every country regression (see Eberhardt 2012 for details on the empirical aspects of estimating heterogenous models using STATA).

First (Maddala and Wu 1999, not reported) and second generation (Pesaran 2007) panel unit root tests applied to output and input data used in this study suggest that nonstationarity cannot be ruled out in this dataset. As in Eberhardt and Teal (2013) our results show strong evidence of cross-section dependence within the full sample dataset, based on the Pesaran (2004) CD test.¹

Heterogeneous parameter models seem to perform better than the traditional pooled models with the neighbor CMG and the AMG2 showing best performance. These models reject nonstationarity, show no evidence of cross-section dependence and do not reject CRS. Table 11.1 present results for these two models and the best performing pooled models (distance and neighbor CCE) compared to estimates of the same models with CRS imposed. The AMG2 model performs better than all other models when CRS are imposed with no significant changes in coefficient values. In contrast, the coefficient for labor in the CCEPd and the CMGn model doubles and other coefficients also change significantly when CRS are imposed. This could be explained in part by the fact that even though these models do not reject CRS or reject it at the 10 % level, the obtained labor coefficients are relatively large compared to the one obtained with the AMG2 model. Imposition of CRS in these models led to larger magnitudes for either the land or the implied labor coefficient.

Input elasticities of our preferred model are not comparable with estimates from other studies using Cobb-Douglas production functions given that most of those studies used five inputs instead of six as they did not include animal feed, which in our results shows a high and significant coefficient. Previous studies also use number of tractors and animal stock as proxies for crop and livestock capital instead of the newly available capital estimates from FAO included in this study. In this context, our results show higher coefficients on crop capital than those obtained by other studies and relatively low coefficients for livestock capital which often shows high coefficients across most specifications.

11.4 TFP Growth and Performance of Sub-Saharan Africa's Agriculture, 1971–2011

Results of the growth decomposition analysis for a sample of 38 Sub-Saharan African countries shows that annual growth per worker for the period 1971–2011 was 0.3 %, or equivalently that SSA's agricultural output per worker was 15 % higher in 2011 than its level in 1971. Two periods with contrasting results can be distinguished in Fig. 11.2 and Table 11.2. A first period of poor performance and decline stretches from the beginning of the period to the mid-1980s, during which growth in SSA was negative: −0.8 during 1971–1980 and −0.1 % from 1981 to 1990. This period is followed by a period of recovery and improved performance

¹Results of estimates and tests for all models are available from authors upon request.

Table 11.1 Best performing models, unrestricted and with CRS imposed

	CCEPn	CCEPd	CMGn	AMG2
	Unrestricted	CRS-imposed	Unrestricted	CRS-imposed
Labor	-0.106 (0.218)	-0.241* (0.135)	-0.203 (0.138)	-0.0389 (0.124)
Crop capital	0.0829*** (0.0283)	0.0863*** (0.0282)	0.124** (0.0475)	0.143*** (0.0513)
Livestock capital	0.229*** (0.0651)	0.263*** (0.0590)	0.307*** (0.0386)	0.196*** (0.0306)
Fertilizer	0.0133 (0.00747)	0.0135* (0.00784)	0.0155*** (0.00504)	0.0131** (0.00494)
Land	0.134 (0.175)	0.206*** (0.0784)	0.134 (0.0953)	0.124 (0.0748)
Feed	0.230*** (0.0538)	0.230*** (0.0516)	0.130*** (0.0276)	0.136*** (0.0282)
Constant	-5.923*** (1.666)	1.722*** (0.233)	-3.520*** (0.819)	-0.943*** (0.0614)
Implied labor coeff.	0.20 0.20	0.08 0.08	0.15 0.15	0.117 0.224

(continued)

Table 11.1 (continued)

	CCEPn	CCEPd	CMGn	AMG2
	Unrestricted	CRS-imposed	Unrestricted	CRS-imposed
Returns ^a	CRS	—	CRS	—
RMSE	0.075	0.078	0.063	0.066
Stationarity ^b	I(0)	I(0)	I(0)	I(0)
Mean ρ_{ij}^c	0.131	0.138	0.146	0.151
CD(p) ^d	2.08	2.23	-1.26	-2.13
CD p value	0.038	0.026	0.207	0.033
Observations	6171	6171	6171	6171
Number of countries	121	121	121	121

Standard errors in parentheses

**p<0.01, **p<0.05, *p<0.1

^aCRS refers to constant returns to scale^bPesaran (2007) CIPS test results: I(0) stationary, I(1) non-stationary^cMean Absolute Correlation coefficient^dPesaran CD test, H0: no cross-section dependence E

Notes: Dependent variable is log output per worker in all models. CCEPn and CCEPd are the common correlated effects pooled estimators using the cross-section averages of the observed output and input variables of contiguous neighbors and cross-section averages calculated using the population weighted geographic distance between countries, respectively. The CMGn model is the heterogeneous version of the CCE model with individual country regressions augmented by cross-section averages of dependent and independent variables using the data for the entire panel, in this case using neighboring countries as weights. The AMG2 is the Augmented Mean Group model with the dynamic process variable imposed in every country regression with a unit coefficient

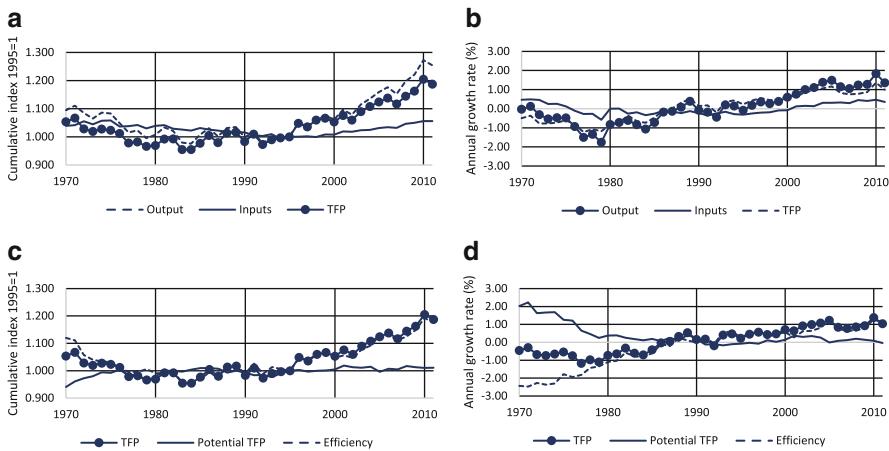


Fig. 11.2 Evolution of levels and growth rates of output per worker and its components, 1971–2011. (a) Output, input and TFP levels. (b) Output, input and TFP growth rates. (c) TFP, potential TFP and efficiency levels. (d) TFP, potential TFP and efficiency growth rates. *Source:* Elaborated by authors

starting in the mid-1990s extending up to 2011, the last year for which information is available. During this period, output per worker grows at an annual rate of 1.5 %, with 0.6 % growth in 1991–2001 and accelerating to 1.5 % in the last decade (2001–2011).

The decomposition of SSA's growth in output per worker into inputs, efficiency and technical change shows that 72 % of growth in output per worker is explained by TFP growth, with inputs explaining the remaining 28 %. Most of the observed TFP growth of the last 15 years is the result of SSA improved efficiency after falling behind during the 1970s and 1980s (last column of Table 11.2). The contribution of technical change to growth after 1995 was insignificant (4 %), occurring only in the last 5 years of the analyzed period. Notice that after reaching the 1971 levels of efficiency in 2005, growth in efficiency decreases to 0.7 % annually compared to 1.5 % in 2001–2005. Figure 11.2 seems to suggest that between 1995 and 2005 the region was catching up to efficiency levels of the early 1970s as it is only in 2005 that SSA reaches the same efficiency levels of 1971. Output per worker recovers to 1971s levels before efficiency as growth in inputs has contributed to output growth since the beginning of the recovery in 1995 at an average rate of 0.4 %.

Among inputs, we observe the expected negative growth of agricultural land related to rapid population growth and a sharp decline in fertilizer use from 4 % growth in the 1970s to negative growth in the 1990s as most countries adjust their economies, and back to increasing fertilizer use in the 2000s. Capital used for livestock production also grew at negative rates during the 1980s and 1990s but started recovering in the 2000s. On the other hand, capital in crop production shows negative growth rates during the whole period (-0.5%). This means that SSA is using at present 20 % less capital per worker in agricultural production

Table 11.2 Growth rates of output and inputs per worker and TFP and its components, different periods

	1971–1980	1981–1990	1991–2000	2001–2011	1971–2011	2001–2005	2006–2011	Growth rate 1995–2011	Contribution to growth ^a 100
Output	-0.83	-0.08	0.61	1.53	0.33	1.78	1.32	1.40	66
Efficiency	-1.22	0.01	0.56	1.04	0.12	1.49	0.67	0.92	
Technical change	0.40	0.14	0.12	0.06	0.18	-0.17	0.26	0.12	9
TFP	-0.83	0.14	0.68	1.10	0.29	1.32	0.92	1.04	74
Inputs	0.00	-0.23	-0.08	0.42	0.04	0.46	0.40	0.36	26
Land	-1.86	-1.85	-1.68	-1.26	-1.66	-1.10	-1.39	-1.33	-48
Crop capital	-0.46	-0.66	-0.44	-0.43	-0.50	-0.76	-0.16	-0.34	-17
Livestock capital	0.22	-0.38	-0.38	0.83	0.09	0.58	1.04	0.61	45
Fertilizer	4.08	1.10	-1.19	2.77	1.70	3.75	1.96	0.93	8
Feed	0.89	1.22	2.14	2.15	1.61	2.81	1.61	2.41	112
Labor ^b	2.16	2.20	2.01	1.79	2.03	1.77	1.82	1.90	n.a.

^aFor inputs contribution is to growth in total inputs^bGrowth in the number of economically active people in agriculture

Source: Elaborated by authors

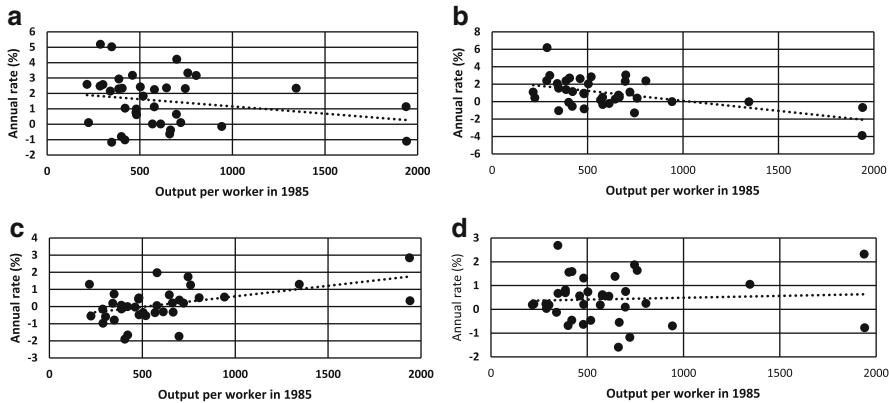


Fig. 11.3 Average annual growth rate between 1995 and 2011 in output per worker, efficiency, technology and inputs against output per worker in 1985. (a) Change in output/worker. (b) Change in efficiency. (c) Change in potential TFP. (d) Change in inputs. *Source:* Elaborated by authors

than it did in 1971. Growth in aggregate input after 1995 was driven by growth in feed and livestock capital. Modest growth in fertilizer and decreases in crop capital and agricultural land per worker indicate that the process of intensification in crop production is driven by labor use with little contribution of inputs and capital.

Figure 11.3 plots average annual growth rates of the different labor productivity-component against output per worker in 1985. Panels A and B, show a significant and negative coefficient between the level of output per worker in 1985 and growth during 1995–2011. The figure suggests that growth after 1995 has mostly benefitted relatively poor countries. The statistically significant positive regression slope coefficient in Panel C indicates that countries with higher output and input per worker in 1985 have benefited much more from technological progress than those countries with low output and input per worker in 1985. This is confirmed by Panel D which shows a positive and statistically significant slope, suggesting that increase in the use of capital and inputs has done little to reduce the gap in labor productivity between countries. In summary, the negative and significant coefficients in panels A and B show that growth after 1995 has been driven by efficiency benefiting low labor productivity (poorer) countries. On the other hand, technical change and increased use of inputs is positively related to higher initial levels of output per worker.

11.5 Productivity Levels and Implications for Growth

After 15 years of agriculture growing at an average rate of 3.3 or 1.5 % per worker, how productive is agriculture in SSA compared to agriculture in other regions and what explains labor productivity differences between SSA and other regions? What effort in terms of inputs and TFP is needed to increase labor productivity? In this section we discuss possible answers to these questions.

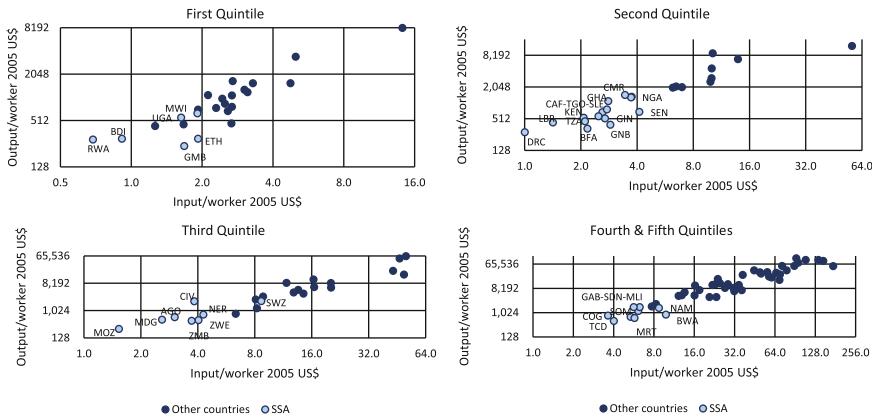


Fig. 11.4 Input and output per worker by quintile of land-labor ratio for SSA countries and other countries, average 2009–2011 (log scale). *RWA* Rwanda, *BDI* Burundi, *UGA* Uganda, *GMB* Gambia, *ETH* Ethiopia, *MWI* Malawi, *DRC* Congo, D.R., *LBR* Liberia, *KEN* Kenya, *TZA* Tanzania, *BFA* Burkina Faso, *GNB* Guinea-Bissau, *GIN* Guinea, *CAF* Central African Rep., *TGO* Togo, *SLE* Sierra Leone, *GHA* Ghana, *SEN* Senegal, *CMR* Cameroon, *NGA* Nigeria, *MOZ* Mozambique, *MDG* Madagascar, *ZMB* Zambia, *ZWE* Zimbabwe, *AGO* Angola, *NER* Niger, *CIV* Cote d'Ivoire, *SWZ* Swaziland, *TCD* Chad, *COG* Congo, Rep., *MRT* Mauritania, *SOM* Somalia, *GAB* Gabon, *MLI* Mali, *SDN* Sudan, *NAM* Namibia, *BWA* Botswana. *Source:* Elaborated by authors

Figure 11.4 plots levels of total input per worker against output per worker for SSA and other countries grouping countries by quintile of land-labor ratio. For example, Fig. 11.4a shows land scarce countries or countries with very low land-labor ratio. Rwanda, Burundi, Uganda, Malawi, Gambia and Ethiopia are SSA's most land-scarce, labor-abundant countries. At the other end, countries with very high land-labor ratio includes the Republic of Congo and Gabon in a tropical-humid agroecology and several arid and semi-arid countries.

The first thing to notice in Fig. 11.4 is the very low level of input per worker used by SSA countries at all levels of land-labor ratio. In Fig. 11.4a, all SSA countries use inputs below 2 dollars per worker when most countries show values between 2 and 4. On average, input per worker in other countries (not including Korea, the country with the highest level of inputs) is twice the average level in SSA countries (2.7 and 1.5 respectively). This difference increases as we move up to countries in higher land-labor quintiles. In the second quintile, use of inputs in SSA countries concentrates around 3 dollars per worker while other countries on average use 8 dollars per worker. Ghana appears as an average SSA country in terms of input use in the second quintile, with Kenya, Tanzania and Burkina Faso on the low side and Nigeria and Cameroon on the high side of input use. In quintile 3, SSA countries average 3 dollars of inputs per worker while average input level in other countries is 13 dollars. Similarly, SSA countries in quintiles 4 and 5 concentrate around values of 5–6 dollars per worker, compared to 32 dollars in South Africa. Also notice in quintiles 1–3 what we can describe as “absolute outliers”, or countries with extremely low levels of input per worker even when compared to other SSA

countries. These are Rwanda and Burundi in the first quintile, D. R. Congo and Liberia in the second quintile and Mozambique (we could possibly add Madagascar and Angola) in the third quintile.

What explains differences in output per worker between countries? A way to summarize the contribution of efficiency (E), factor endowments (F) and available technology or potential TFP (T) to output differences is the variance decomposition. Aggregating inputs, the Cobb-Douglas production function can be expressed as $Y = AF$, where A is TFP and is equal to $A = ET$, or the product of efficiency and available technology. The variance of log output per worker can be decomposed as:

$$\begin{aligned} \text{Var}(\ln Y) &= \text{Var}(\ln T) + \text{Var}(\ln E) + \text{Var}(\ln F) + 2\text{Cov}(T, E) \\ &\quad + 2\text{Cov}(T, F) + 2\text{cov}(E, F) \end{aligned}$$

Table 11.3 presents the contribution of factors, efficiency and technology to the variation of output per worker in agriculture calculated separately for the four groups of countries in different land-labor quintiles. Results show that differences in labor productivity between countries in 2001–2011 are explained in all cases mostly by differences in input levels. The direct effect of inputs is largest in the second quintile, the group with the highest number of SSA countries (76 %) and is higher than 60 % in all other groups. Adding direct and indirect input effects (through technology) we find that inputs explain more than 80 % of labor productivity differences between countries in all groups.

Comparing values in different periods we observe two interesting results. First, the contribution of efficiency to differences in labor productivity has decreased

Table 11.3 Contribution of factors, efficiency and technology to the variation of output per worker in agriculture in different periods, SSA and other countries

	Efficiency	Inputs	Technology	Direct and indirect input effect
<i>First quintile</i>				
1981–1990	0.29	0.62	0.08	0.71
1991–2000	0.26	0.61	0.13	0.74
2000–2011	0.16	0.63	0.20	0.84
<i>Second quintile</i>				
1981–1990	0.17	0.77	0.07	0.83
1991–2000	0.11	0.77	0.12	0.89
2000–2011	0.07	0.76	0.17	0.93
<i>Third quintile</i>				
1981–1990	0.30	0.55	0.14	0.70
1991–2000	0.25	0.58	0.16	0.75
2000–2011	0.21	0.60	0.19	0.79
<i>Fourth and fifth quintiles</i>				
1981–1990	0.21	0.65	0.14	0.79
1991–2000	0.19	0.64	0.17	0.81
2000–2011	0.16	0.65	0.19	0.84

Source: Elaborated by authors

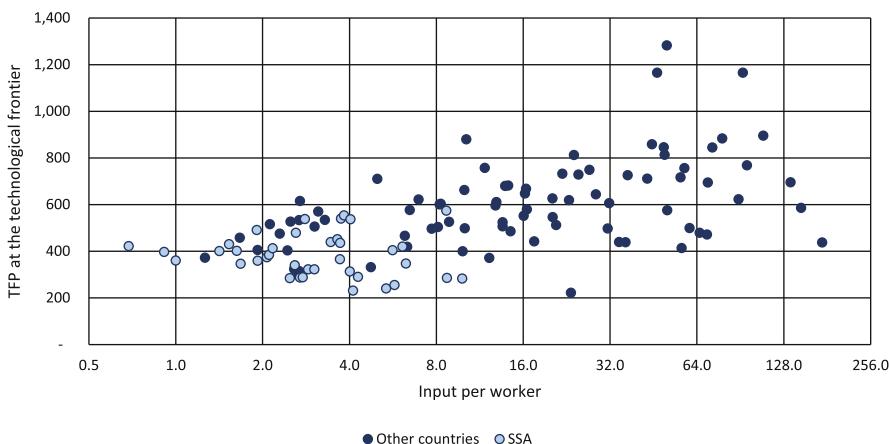


Fig. 11.5 Input per worker and potential TFP levels 2009–2011. *Source:* Elaborated by authors

with time, especially in the first and second quintile which coincides with growth patterns in SSA in recent years driven by efficiency gains, a process to which other labor intensive agricultures, particularly in Asia, have contributed. Second, the importance of inputs explaining labor productivity has increased with time, not because of its direct effect, which has changed very little, but through its indirect effect, appropriate technology. In other words, differences in labor productivity are in part the result of low intensity in the use of inputs per worker but also result from low productivity of the input mix used.

Is the level of the technological frontier affected by the input mix? Figure 11.5 plots input per worker against TFP at the frontier for all countries in our sample. The figure suggests an “appropriate technology” or some level of input per worker that results in higher productivity levels. Most SSA countries produce with levels of input per worker below 4 dollars where we observe the lowest frontier levels (400 dollars per unit of input). Countries using 10 dollars of input per worker can potentially produce 530 dollars per unit of input, meaning that these countries have access to technologies that can produce approximately 30 % more output per unit of input than technologies that SSA countries can use. Countries using the most productive input mix can produce 75 % more output per unit of input than SSA countries.

The difference in TFP at different levels of input per worker that we observe in Fig. 11.5 can be tested more rigorously using Analysis of Variance (ANOVA). To do this we classify countries in ten quantiles by level of input per worker. We form one group with the first four quantiles including most SSA countries and countries using inputs below 4 dollars per worker. We then compare average TFP levels and TFP changes in time of this group against all other groups. Results are shown in Table 11.4.

Results show that technologies available to countries using low levels of input per worker (4 dollars or less) are significantly less productive than those technologies available to countries using at least 10 dollars of input per worker. For example,

Table 11.4 ANOVA comparing TFP levels and shifts in the frontier between 1971–1980 and 2001–2011 at different levels of input/worker relative to shifts in the first four quantiles of input per worker

	Input/worker	Output/worker	Diff. in TFP levels ^a	P > t	Diff. in TFP change ^b	P > t
q1–q4	2.7	409	0	–	0.0	–
q5	6.7	439	30	0.031	14.3	0.069
q6	10	527	118	0.000	10.0	0.199
q7	15	530	121	0.000	17.4	0.027
q8	24.7	573	164	0.000	15.7	0.046
q9	45.7	718	309	0.000	28.8	0.000
q10	95.3	687	278	0.000	35.7	0.000

^aNumber of obs. = 1320, Prob. >F = 0.000, R-squared = 0.38, Adj R-squared = 0.38

^bNumber of obs. = 120, Prob. >F = 0.000, R-squared = 0.21, Adj R-squared = 0.17

Source: Elaborated by authors

countries using 15 dollars of input per worker show TFP levels at the frontier of 530 dollars compared to 409 for countries using 2.7 dollars of input per worker, while the differences with the most productive group is 278 dollars, all differences significant at the 0.1 % level.

Results also show that technical change between 1971–1980 and 2001–2011 shifted the frontier unevenly with TFP in those portions of the frontier at low levels of input per worker (where SSA countries are located). The speed of technical change at low levels of output per worker is significantly lower than technical change occurring at input per worker of 15 dollars and much lower than that occurring at 30 dollars or higher of input per worker. These results suggest that technological divergence is taking place in agriculture, increasing the distance between countries with the “right” input mix and countries (like SSA countries) producing at very low levels of input per worker.

11.6 Conclusions

We revisited past performance of agriculture in SSA and found that improved technical efficiency has been the main driver of growth in recent years benefiting poorer, low labor productivity countries. On the other hand, we observed that countries with higher output and input per worker have benefited much more from technological progress than poorer countries suggesting that technical change has done little to reduce the gap in labor productivity between countries.

A possible explanation of these results can be found in a literature that has emphasized the potential dependence of productivity on inputs to explain differences in income levels and the lack of convergence in labor productivity. Under this approach, the technological frontier is not the same for all countries as some technologies may be more or less productive than others depending on the country’s relative input mix. This is because advanced countries invent technologies that are

compatible with their own factor mix, but these technologies are less productive when used with the very different factor mix of poor countries. To get a better understanding of the role of inputs on TFP gaps, this study used a growth accounting approach to analyze the explanatory power of the appropriate technology hypothesis to explain differences in productivity levels between SSA and other countries.

Our findings show that the levels of input per worker used in SSA's agriculture at present are extremely low and that differences in labor productivity among 121 high income and developing countries are explained mostly by differences in the intensity in the use of inputs. We also found that the importance of inputs explaining labor productivity has increased with time, not because of its direct effect which has changed very little, but because of a growing gap in the productivity of inputs as the result of low productivity of the input mix in poor countries. Countries using the most productive input mix can produce 75 % more output per unit of input than SSA countries. These differences in TFP can increase in the future as we found that technical change shifts the frontier unevenly, with TFP in those portions of the frontier at higher levels of input per worker growing much faster than those portions at low levels of input per worker (where SSA countries are located). A possible interpretation of the growing importance of technology explaining differences in output per worker is that technological divergence is taking place in agriculture, increasing the distance between countries with the “right” input mix and countries (like SSA countries) producing at very low levels of input per worker.

The existence of an appropriate technology could have significant implications for policy and development. Is the slow pace of technology adoption and TFP growth in SSA the result of inappropriateness of technology given the very particular conditions and low levels of capitalization of agriculture in these countries? Should countries adapt technologies produced by advanced countries to their own input mix, or should they “adapt” their agricultural sector to use modern technologies more efficiently (e.g., the poor smallholders vs. commercial agriculture debate)? The appropriate technology hypothesis could bring a different perspective to this debate already taking place in SSA.

Acknowledgments This work was undertaken as part of the Agricultural Science and Technology Indicators (ASTI) and the CGIAR Research Program on Policies, Institutions, and Markets (PIM) led by the International Food Policy Research Institute (IFPRI). Funding support for this study was provided by the Bill & Melinda Gates Foundation, the Canada Department of Foreign Affairs, and PIM. I thank Markus Eberhardt for sharing STATA code he developed and I adapted and used in the econometric analysis of this study. Any errors are my own responsibility. The opinions expressed here belong to the author, and do not necessarily reflect those of PIM, IFPRI, or CGIAR.

Appendix

Countries in our sample were grouped by agroecological zone (AEZ). These zones were defined based on data from Lee et al. (2005). Table 11.5 presents the country classification by AEZ.

Table 11.5 Classification of SSA countries by agroecology

Latitude	Length of growing period (days)	LGP range	LGP class	Country
Tropical or temperate	50	<100	Arid	Botswana
	95	<100		Chad
	78	<100		Mali
	38	<100		Mauritania
	54	<100		Namibia
	49	<100		Niger
	37	<100		Somalia
Temperate	186	100–200	Semi-arid	Swaziland
	132	100–200		Zimbabwe
Tropical	270	>260	Humid	Burundi
	276	>260		Congo, D.R.
	328	>260		Congo, Rep.
	350	>260		Eq. Guinea
	325	>260		Gabon
	277	>260		Cote d'Ivoire
	339	>260		Liberia
	312	>260		Rwanda
	303	>260		Uganda
	173	100–260	Semi-arid and Sub-humid	Angola
	201	100–260		Benin
	140	100–260		Burkina
	255	100–260		Cameroon
	232	100–260		Central Afr. Rep.
	148	100–260		Ethiopia
	150	100–260		Gambia
	258	100–260		Ghana
	232	100–260		Guinea
	180	100–260		Guinea-Bissau
	129	100–260		Kenya
	237	100–260		Madagascar
	176	100–260		Malawi
	184	100–260		Mozambique
	165	100–260		Nigeria
	107	100–260		Senegal
	250	100–260		Sierra Leone
	115	100–260		Sudan
	199	100–260		Tanzania
	230	100–260		Togo
	173	100–260		Zambia

Source: Elaborated by authors based on Lee et al. (2005)

References

- Acemoglu D, Zilibotti F (2001) Productivity differences. *Q J Econ* 116(2):563–606
- Alene AD (2010) Productivity growth and the effects of R&D in African agriculture. *Agric Econ* 41(3-4):223–238
- Arnade C (1998) Using a programming approach to measure international agricultural efficiency and productivity. *J Agric Econ* 49:67–84
- Barro RJ, Sala-i-Martin X (1991) Convergence across states and regions. *Brook Pap Econ Act* 1991(1):107–182
- Basu S, Weil DN (1998) Appropriate technology and growth. *Q J Econ* 113(4):1025–54
- Block SA (1995) The recovery of agricultural productivity in Sub-Saharan Africa. *Food Policy* 20:385–405
- Block S (2010) The decline and rise of agricultural productivity in sub-Saharan Africa since 1961 (No. w16481). National Bureau of Economic Research, Cambridge
- Bravo-Ortega C, Lederman D (2004) Agricultural productivity and its determinants: revisiting international experiences. *Estud Econ* 31(2):133–163
- Bureau C, Färe R, Grosskopf S (1995) A comparison of three nonparametric measures of productivity growth in European and United States Agriculture. *J Agric Econ* 45:309–26
- Caves DW, Christensen LR, Diewert WE (1982) The economic theory of index numbers and the measurement of input, output, and productivity. *Econometrica* 50:1393–414
- Cermenio R, Maddala G, Trueblood M (2003) Modelling technology as a dynamic error components process: the case of the inter-country agricultural production function. *Econ Rev* 22:289–306
- Chavas JP (2001) An international analysis of agricultural productivity. In: Zepeda L (ed) Agricultural investment and productivity in developing countries. Food and Agriculture Organization, Rome
- Coakley JA, Fuertes M, Smith RP (2006) Unobserved heterogeneity in panel time series models. *Comput Stat Data Anal* 50(9):2361–2380
- Craig BJ, Pardey PG, Roseboom J (1997) International productivity patterns: accounting for input quality, infrastructure, and research. *Am J Agric Econ* 79(4):1064–1076
- Eberhardt M (2012) Estimating panel time-series models with heterogeneous slopes. *Stata J* 12(1):61–71
- Eberhardt M, Bond S (2009) Cross-section dependence in nonstationary panel models: a novel estimator. MPRA Paper 17692 University Library of Munich. http://mpra.ub.uni-muenchen.de/17692/1/MPRA_paper_17692.pdf. Accessed 15 February 2014
- Eberhardt M, Teal F (2013) No mangoes in the tundra: spatial heterogeneity in agricultural productivity analysis. *Oxf Bull Econ Stat* 75:914–939
- FAO (Food and Agriculture Organization of the United Nations) (2014) FAOSTAT database. <http://www.fao.org/>. Accessed 20 Jan 2014
- Fuglie KO (2011) Agricultural productivity in Sub-Saharan Africa. In: Lee DL (ed) The food and financial crisis in Africa. Commonwealth Agricultural Bureau International, Wallingford
- Fuglie KO, Rada N (2012) Constraints to raising agricultural productivity in Sub-Saharan Africa. In: Fuglie KO, Wang SL, Ball VE (eds) Productivity growth in agriculture: an international perspective. Commonwealth Agricultural Bureau International, Wallingford
- Fulginiti L, Perrin RK (1997) LDC agriculture: nonparametric Malmquist productivity indexes. *J Dev Econ* 53:373–90
- Fulginiti L, Perrin RK (1999) Have price policies damaged LDC agricultural productivity? *Contemp Econ Policy* 17:469–75
- Fulginiti LE, Perrin RK, Yu B (2004) Institutions and agricultural productivity in Sub-Saharan Africa. *Agric Econ* 4:169–80
- Griliches Z (1964) Research expenditures, education, and the aggregate agricultural production function. *Am Econ Rev* 6:961–74
- Grossman G, Helpman E (1991) Innovation and growth in the global economy. MIT Press, Cambridge

- Growiec J (2012) The world technology frontier: what can we learn from the US states? *Oxf Bull Econ Stat* 74(6):777–807
- Hayami Y, Ruttan V (1985) Agricultural development: an international perspective. Johns Hopkins University Press, Baltimore
- Jaffe A (1986) Technological opportunity and spillovers of R&D: evidence from firms' patents, profits and market value. *Am Econ Rev* 76(5):984–1001, NBER Working Paper Series No. 1815
- Jerzmanowski M (2007) Total factor productivity differences: appropriate technology vs. efficiency. *Eur Econ Rev* 51:2080–2110
- Kumar S, Russell RR (2002) Technological change, technological catch-up, and capital deepening: relative contributions to growth and convergence. *Am Econ Rev* 92(3):527–548
- Lee HL, Hertel TH, Sohngen B, Ramankutty R (2005) Towards an integrated land use data base for assessing the potential for greenhouse gas mitigation. GTAP Technical Papers, 26
- Ludena CE, Hertel TW, Preckel PV, Foster K, Nin A (2007) Productivity growth and convergence in crop, ruminant, and non-ruminant production: measurement and forecasts. *Agric Econ* 37(1):1–17
- Lusigi A, Thirtle C (1997) Total factor productivity and the effects of R&D in African agriculture. *J Int Dev* 9:529–38
- Maddala GS, Wu S (1999) A comparative study of unit root tests with panel data and a new simple test. *Oxf Bull Econ Stat* 61(Special Issue):631–652
- Nin A, Arndt C, Preckel PV (2003) Is agricultural productivity in developing countries really shrinking? New evidence using a modified non-parametric approach. *J Dev Econ* 71:395–415
- Nin-Pratt A, Yu B (2012) Agricultural productivity and policy changes in Sub-Saharan Africa. In: Fuglie KO, Wang SL, Ball VE (eds) Productivity growth in agriculture: an international perspective. Commonwealth Agricultural Bureau International, Wallingford
- Parente SL, Prescott EC (1994) Barriers to technology adoption and development. *J Polit Econ* 102(2):298–321
- Pesaran MH (2004) General diagnostic tests for cross section dependence in panels. CESifo Working Paper 1229; IZA Discussion Paper 1240. University of Cambridge
- Pesaran MH (2006) Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* 74(4):967–1012
- Pesaran MH (2007) A simple panel unit root test in the presence of cross-section dependence. *J Appl Econ* 22(2):265–312
- Pesaran MH, Smith RP (1995) Estimating long-run relationships from dynamic heterogeneous panels. *J Econ* 68(1):79–113
- Prasada Rao DS, Coelli TJ (1998) Catch-up and convergence in global agricultural productivity, 1980–1995. CEPA Working Papers 4/98, Department of Econometrics, University of New England, Armidale
- Rambaldi AN, Prasada Rao DS, Dolan D (2007) Measuring productivity growth performance using meta-frontiers with applications to regional productivity growth analysis in a global context. Australasian Meeting of the Econometric Society, 3–6 July. Brisbane
- Segerstrom PS, Anant TCA, Dinopoulos E (1990) A Schumpeterian model of the product life cycle. *Am Econ Rev* 80:1077–1092
- Suhariyanto K, Thirtle C (2001) Asian agricultural productivity and convergence. *J Agric Econ* 52:96–110
- Suhariyanto K, Lusigi L, Thirtle C (2001) Productivity growth and convergence in Asian and African agriculture. In: Lawrence P, Thirtle C (eds) Asia and Africa in comparative economic perspective. Palgrave, London
- Trueblood MA, Coggins J (2003) Intercountry agricultural efficiency and productivity: a Malmquist index approach. Mimeo. World Bank, Washington, DC
- USDA (United States Department of Agriculture) (2014). Economic research service: dataset on international agricultural productivity. <http://www.ers.usda.gov/data-products/international-agricultural-productivity.aspx>. Accessed 12 Aug 2014

Chapter 12

University Knowledge Spillovers and Innovative Startup Firms

Leonard Sabetti

Abstract In this paper, we investigate the role of university knowledge spillovers in fostering innovative startup firms, measured by R&D intensity, an important predictor of firm innovation and productivity. We use annual data from the Kauffman Firm Survey of a representative cohort of U.S. startups over the period 2004–2011. Controlling for individual-firm characteristics and local factors, we test the effects of regional variation in R&D intensity of the higher-education sector on startup firms' R&D expenditure decisions. We find strong effects on both extensive and intensive margins of firm R&D expenditures. Our results shed light on the role of entrepreneurs and new firm formation as a mechanism for innovation and universities as an important source of knowledge and technology transfer.

Keywords Innovation • Productivity • Knowledge spillovers • Entrepreneurship

12.1 Introduction

The advent of rich micro-level data at both firm and worker levels has increasingly uncovered the dynamic, Schumpeterian nature of the U.S. economy unmatched by any other. A main theoretical prediction of the firm dynamics and entrepreneurship literature emphasizes that mature, large firms tend to stagnate whereas new, young firms tend to grow very quickly conditional on survival. Haltiwanger et al. (2010) finds that a significant contribution of both gross and net job creation stems from the launch of new business ventures. However, the majority of new firms fail within 10 years or remain small while a fraction exhibit disproportionately high growth rates, contributing to the bulk of productivity and job gains. Understanding the systemic factors (or barriers) to the creation of innovative, young firms remains an important public policy concern.

L. Sabetti (✉)

George Mason University, 4400 University Dr., Fairfax, VA 22030, USA

e-mail: lsabetti@gmu.edu

Recently, the entrepreneurship literature has emphasized the institutional context of the local area as an important factor in explaining the type of firms that are launched and their performance over time. For instance, the knowledge spillover theory of entrepreneurship (KSTE), set forth by Acs et al. (2013), views the entrepreneur as a conduit for innovation by commercializing ideas that evolved from an incumbent organization via the creation of a new firm. According to this view, it is the act of launching a firm, or the entry margin, whereby gains in productivity or innovation will occur. A local area rich in knowledge generates the opportunities from those ideas, or increases the chances that successful ventures will be brought to fruition by motivated and alert entrepreneurs (Kirzner 1997; Shane 2000). Boettke and Coyne (2009) emphasize how the local institutional environment encourages (or not) entrepreneurs to bet on their ideas and obtain financing to bring those ideas to life.

Under the view that highly innovative startups are a result of their surroundings, we set out to investigate the extent to which the regional knowledge base drives innovation among new firms using a novel data set for the United States. As a result, we contribute to a theory on high-growth innovative firms. We study a cohort of startup firms launched in 2004, with data collected annually up until 2011. Our contribution to the literature stems from the key features of our data. First, the data contains detailed micro-level information that are not usually featured in administrative firm-level data. Second, the data is the first of its kind to survey a comprehensive and nationally representative sample of firms beginning from their time of origin and that over-samples firms in the high-tech sector. Third, given the recent period which the data covers, we are able to say something about to what extent knowledge spillovers are still relevant at the local level in a world of increasing inter-connectivity.

To construct a measure of the regional knowledge base, we turn to one of the world's oldest institutions: the university. Adams (2001), Hall et al. (2003), Siegel and Phan (2004) and Huynh and Rotondi (2009) document the influence of university research as well as university-industry cooperation as an important channel for promoting firm R&D expenditures. The geography of university-industry cooperation may also play an important role given evidence that knowledge spillovers are mostly localized, for example Jaffe et al. (1993). Little empirical evidence exists on the linkages between knowledge spillovers stemming from higher education sector to small, young startup firms.

Section 12.2 describes the data and provides an overview of the empirical strategy. Section 12.3 presents the results and Sect. 12.4 concludes.

12.2 Data and Empirical Methodology

The Ewing Marion Kauffman Foundation commissioned the largest to date longitudinal study of new businesses in the US, known as the Kauffman firm survey (KFS). The survey follows 4,928 firms that launched in 2004 annually until 2011.

The survey questionnaire contains detailed information on the firm, including industry, physical location, employment, profits, intellectual property, business strategy, and financial capital, as well as information on business owners, including age, gender, race, ethnicity, education, previous industry experience, and previous start-up experience. Importantly, the initial survey design called for 5,000 interviews, with a target of 3,000 interviews for high-technology businesses given particular interest in these firms among researchers.

The sampling frame for the panel of business startups was created using the Dun & Bradstreet (D&B) database of business establishments that started in 2004 in the United States, which totaled roughly two hundred and fifty-thousand firms. D&B maintains a large commercial database of businesses compiled through various public and industry sources. In order to obtain a larger sample of startups in high-technology fields, the data was partitioned into strata according to industrial technology categories, based on a classification scheme developed by the Bureau of Labor Statistics, see Hadlock et al. (1991). The final classifications of high and medium technology businesses were determined according to each industry's respective share of employment in research and development (R&D) using data from the BLS Occupational Employment Statistics program and based on three-digit level standard industry classification (SIC) code.

The National science foundation (NSF) Survey of Research and Development Expenditures at Universities and Colleges (academic R&D expenditures survey) is the primary source of information on separately budgeted research and development (R&D) expenditures by academic institutions in the United States and outlying areas and is publicly available.¹ The data provide the precise location for each university which allow for the construction of aggregate spending measures by geographic area.

Regions with a strong presence of research intensive universities may also be associated with strong industrial clusters which transmit knowledge of their own. For instance, Jaffe finds that patent inventors are often likely to cite patents whose authors are located nearby which also may point to the presence of industrial clusters. To control for this, we use a measure of patent application concentration by economic area. The data is drawn from OECD's Regional Patent Database which contains information on patent applications over the period 1978–2008, including patent applications originating from the U.S. as well as the patent holder's geographical location, see Braymen et al. (2011) for more details.

Firms invest in R&D as a means to innovate and improve productivity, often building upon other institutions' R&D efforts. Understanding a firm's R&D expenditures entails two decisions: the extensive margin, or the decision to invest in R&D, and the intensive margin, how much. However, we only observe an expenditure amount for firms that chose to engage in R&D, raising a type of sample selection issue. To account for this using a Heckman selection model one would need to observe factors that affect the participation decision but not the outcome decision.

¹<http://www.nsf.gov/statistics/srvyrdexpenditures/>.

Alternatively, using sample weights in the R&D expenditure equation may mitigate this issue. In our baseline approach, we model separately the extensive and intensive margins of R&D investment.

In the equations we estimate, we control for firm-specific characteristics as well as local factors by area such as real GDP and a proxy for industrial knowledge spillovers. Firm characteristics include features of the business (industry dummies, business location, legal status etc.) and initial financing conditions, such as startup capital (debt + equity), and leverage. While the decision to invest in R&D changes over time, we treat the dependent variable as the firm's average annual R&D investment and estimate the model cross-sectionally.

A crucial issue in trying to assess the effect of university research on the R&D decisions of entrepreneurs is the endogeneity between the startup firm's choice of location and its proximity to university research centers. Do regions with increased knowledge spillovers stemming from R&D investment in the higher education sector lead to more R&D intensive startup firms? Or are entrepreneurs with innovative ideas that require R&D investment more likely to choose their startup location based on the presence of research intensive universities or centers so that they may benefit from access to exiting knowledge and human capital? While we cannot rule out the possibility that the startup location was endogenously chosen, the literature has found evidence of a "home bias"—entrepreneurs tend to launch their business in proximity to their home.

We propose an instrumental variable (IV) approach to provide additional robustness check on the causality of our estimates. Our baseline regressions employ a time-invariant local knowledge variable that is measured as of 2003 to ensure it is pre-determined relative to when the firms in our sample begin operations in 2004. We make use of the Bayh-Doyle Act of 1980 as an exogenous shock to universities' relationship with industry. The Act granted universities full commercial rights over intellectual property developed with public sector funds, thereby incentivizing universities to partner with industry and seek out research projects that may lead to commercially viable inventions. Prior to the Act, these incentives were arguably not as apparent therefore limiting the extent to which the private sector affected university research. To the extent that the shift in policy was enacted independent of any other related factors suggests using the local area's presence of research intensive universities pre-Bayh Doyle as a natural instrumental variable for our local knowledge variable pre-2004. For example, Mowery et al. (2001) has stated: "the principal positive effect of Bayh-Doyle probably has been to motivate universities to reach out more to industry."

We also run our estimates splitting the sample according to whether the founder is a serial entrepreneur. We hypothesize that serial entrepreneurs, which have founded one or more previous businesses, are potentially more likely to strategically choose their startup location. If the effect of university knowledge spillovers is more pronounced for this later group, then this would provide evidence that our estimates suffer from endogeneity.

12.3 Results

In our baseline model, we consider the firm's average annual R&D investment in deflated constant dollars as the dependent variable. Table 12.1, column 1, presents our key findings of the determinants of the intensive margin of R&D. In accordance with theory, credit risk is highly significant and negatively impacts R&D expenditures, possibly measuring a firm's level of financial constraint as in Kaplan and Zingales (2000). Debt to equity ratios, or leverage, also matter as a firm with a low debt to equity ratio spend roughly 75 % more on R&D relative to firms with no debt. High leverage ratios tend to be correlated with a firm that is growing rapidly, but also may suggest a "shadow of death", or in other words a firm on verge of bankruptcy, as in Huynh et al. (2010). Zingales finds that highly leveraged firms also have difficulty in undertaking investments in the future. The literature also finds that firms prefer to finance R&D spending with internal funds, or out of retained earnings. Total startup financial capital is negatively associated with R&D expenditures but the effect is weak. One interpretation is that total startup financial capital is a proxy for startup firm size and that more R&D intensive firms are in fact smaller at their founding. However, this is contradicted by the fact that the number of startup employees, another indicator which is a good proxy for firm size, is strongly associated with increased expenditures in R&D.

We find a treatment effect of roughly 27 % for firms which report direct ties to a university in the survey. In column 2, we base our estimates on the sample of firms that do not partner with universities, so as to obtain a measure of an indirect effect stemming from university knowledge spillovers. We obtain a marginal effect of 0.065 which is significant at the 10 % level. For robustness, we instrument our measure of university knowledge spillovers, based on the average annual R&D expenditure of the local HES from 1999 to 2004, using the 5-year average pre-Bayh Doyle, 1976–1980. The IV estimate is qualitatively similar. The results suggest that a 10 % increase in local HES R&D expenditures translates to a 0.6 % increase in the startup firm's R&D expenditures on average (both variables are log-transformed).

Column three produces the estimates for the sample of entrepreneurs who have not founded any previous business. The estimate for the effect of university knowledge spillovers is quantitatively similar to the baseline results whereas the effect for the sample of serial entrepreneurs is not statistically different from zero, providing some evidence that endogeneity is not biasing our estimates.

Finally, we split the sample according to whether firms are in the medium/high-tech sectors or not. We find that there is no effect for firms not in the high or medium-tech sector, while the effect for this later group rises to roughly 0.102. This result coincided with our initial hypothesis that the regional knowledge base would not affect firms in the non-high tech sector. This result also provides some evidence that our estimates are not driven by confounding factors.

Table 12.1 Intensive Margin: R&D

	Direct effect	Indirect effect (firms w/no univ ties)			
		All	No previous startups	Non-tech	High-tech
Credit risk	−0.248*** 0.08	−0.268*** 0.09	−0.179 0.14	0.034 0.11	−0.516*** 0.13
Log startup financial capital	−0.170** 0.08	−0.167* 0.09	−0.046 0.16	−0.109 0.13	−0.195* 0.1
Squared	0.022*** 0	0.023*** 0.01	0.019** 0.01	0.021*** 0.01	0.024*** 0.01
Debt/equity < 0.5	−0.377** 0.17	−0.414** 0.18	−0.079 0.26	−0.185 0.26	−0.588*** 0.22
0.5 ≤ debt/equity < 1.0	−0.487** 0.21	−0.334 0.21	−0.036 0.33	−0.028 0.31	−0.519 0.32
1.0 ≤ debt/equity < 2.0	−0.588*** 0.21	−0.676*** 0.21	−0.481 0.3	−0.216 0.29	−0.941*** 0.25
2.0 ≤ debt/equity	−0.756*** 0.19	−0.694*** 0.22	−0.550* 0.33	−0.503* 0.27	−0.731** 0.29
Revenues in startup year	−0.359*** 0.12	−0.217* 0.13	−0.159 0.19	−0.19 0.21	−0.228 0.16
Log startup # employees	0.167*** 0.06	0.174*** 0.06	0.245*** 0.08	0.068 0.1	0.273*** 0.09
High tech	0.482*** 0.15	0.356** 0.15	0.527*** 0.17		
Patent applications index		2.674 1.71	1.847 1.59	2.14 1.9	3.600** 1.78
Univ partnership	0.270* 0.16				
Local HES R&D expenditures		0.065** 0.03	0.069* 0.04	0.044 0.04	0.102** 0.04
IV estimate		0.058* 0.03			0.120** 0.05
N	838	716	400	328	388

Note: *, **, *** denotes statistical significance at the 10, 5, and 1 % levels respectively. Based on active firms in 2007, 2008, 2009 and 2010. Clustered standard errors by economic area. Dependent variable is Log R&D expenditures conditional on positive amount. Startup financial capital is log of total debt and equity in baseline year. Base category for leverage ratios is No Debt. Local HES Knowledge Spillover is the log average annual R&D expenditure amount from the HES in the economic area for 1999–2004. Includes industry fixed effects, business characteristics and owner demographics

12.4 Conclusion

Using the Kauffman Firm Survey, a novel dataset on startup firms launched in the U.S. in 2004, we find evidence that the regional knowledge base fosters innovation among new business ventures. Our results hold after conducting a series of robustness checks. Our results highlight the importance of new firms as a source of innovation and universities as important drivers of knowledge transfer.

Acknowledgements The author “Leonard Sabetti” is grateful for financial support from the Ewing Marion Kauffman Foundation and the SSHRC through the Network to Study Productivity in Canada from a Firm-Level Perspective.

References

- Acs Z, Audretsch D, Lehmann E (2013) The knowledge spillover theory of entrepreneurship. *Small Bus Econ* 41(4):757–774
- Adams J (2001) Comparative localization of academic and industrial spillovers. Discussion Paper WP8292, NBER
- Boettke PJ, Coyne CJ (2009) Context matters: institutions and entrepreneurship. *Found Trends(R) Entrep* 5(3):135–209
- Braymen C, Briggs K, Boulware J (2011) R&D and the export decision of new firms. *South Econ J* 78(1):191–210
- Hadlock P, Heckler D, Gannon J (1991) High-technology employment: another view. *Mon Labor Rev* 114(7):26–30
- Hall BH, Link AN, Scott JT (2003) Universities as research partners. *Rev Econ Stat* 85(2):485–491
- Haltiwanger J, Jarmin RS, Miranda J (2010) Who creates jobs? Small vs. large vs. young. Working papers 10–17, Center for Economic Studies, U.S. Census Bureau
- Huynh KP, Petrunia RJ, Voia M (2010) The impact of initial financial state on firm duration across entry cohorts-super-*. *J Ind Econ* 58(3):661–689
- Huynh KP, Rotondi Z (2009) R&D spending and knowledge spillovers. Manuscript
- Jaffe AB, Trajtenberg M, Henderson R (1993) Geographic localization of knowledge spillovers as evidenced by patent citations. *Q J Econ* 108(3):577–598
- Kaplan SN, Zingales L (2000) Investment-cash flow sensitivities are not valid measures of financing constraints. *Q J Econ* 115(2):707–712
- Kirzner IM (1997) Entrepreneurial discovery and the competitive market process: an Austrian approach. *J Econ Lit* 35(1):60–85
- Mowery DC, Nelson RR, Sampat BN, Ziedonis AA (2001) The growth of patenting and licensing by U.S. universities: an assessment of the effects of the Bayh-Dole act of 1980. *Res Policy* 30(1):99–119
- Shane S (2000) Prior knowledge and the discovery of entrepreneurial opportunities. *Organ Sci* 11(4):448–469
- Siegel DS, Phan PH (2004) Analyzing the effectiveness of university technology transfer: implications for entrepreneurship education. Rensselaer working papers in economics 0426, Rensselaer Polytechnic Institute, Department of Economics

Chapter 13

Accounting for Natural Capital in Productivity of the Mining and Oil and Gas Sector

Pat Adams and Weimin Wang

Abstract This paper presents a growth accounting framework in which subsoil mineral and energy resources are recognized as natural capital input into the production process in two ways. Firstly, the income attributable to subsoil resources, or resource rent, is estimated as a surplus value after all extraction costs and normal returns on produced capital have been accounted for. The value of a resource reserve is then estimated as the present value of the future resource rents generated from the efficient extraction of the reserve. Secondly, with extraction as the observed service flows of natural capital, multifactor productivity growth and sources of economic growth can be reassessed by updating income shares of all inputs and then by estimating the contribution to growth coming from changes in the value of natural capital input.

The empirical results on the Canadian oil and gas extraction show that, adding natural capital increases the annual multifactor productivity growth in the oil and gas sector from -2.2 to -1.5 % over the 1981–2009 period. During the same period, the annual real value-added growth in this industry was 2.3 %, of which about 0.3 percentage points or 15 % comes from natural capital.

Keywords Natural resource • Natural capital • Resource rent • Multifactor productivity

JEL code: O40, Q30

P. Adams

Enterprise Statistics, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa,
ON K1A 0T6, Canada

W. Wang (✉)

Economic Analysis Division, Statistics Canada, 100 Tunney's Pasture Driveway,
Ottawa, ON K1A 0T6, Canada

e-mail: Weimin.Wang@statcan.gc.ca

13.1 Introduction

This paper has two objectives. The first is to estimate the resource rent generated through the extraction of subsoil mineral and energy resources and the associated monetary value of resource reserves¹ in Canadian mining industries—what is referred to here as the value of natural capital. The second is to treat subsoil resources themselves as a factor input in resource extractions. This is done by estimating the flow of services derived from natural capital input, and adding it to the value of labour and produced capital inputs in the standard multifactor productivity estimating equation. This produces a measure of multifactor productivity growth that is more complete and provides an estimate of the significance of subsoil resources as a source of economic and productivity growth in the Canadian mineral and resource sector.

Subsoil mineral and energy resources are treated as non-produced and non-financial assets in the System of National Accounts (SNA). To be consistent, exploration and development expenditures are capitalized as produced capital assets in SNA. Therefore, the value of subsoil resources as non-produced assets reflects only the value of resource scarcity.

The productivity accounts calculates multifactor productivity (MFP) growth as the difference between the growth in output and a weighted average growth of all inputs—one of which is the capital derived from investments in fixed assets. The fixed assets included in the accounts for the mining, oil and gas industries include investments in machinery and equipment, structures, and engineering assets such as mine shafts, as well as exploration and development expenditures. Natural capital—the value of the resources is not included.

This paper offers a way in which this can be done and provides estimates of MFP growth when the cost of using natural capital is included. Specifically, the resource rent of subsoil assets is calculated as a surplus value after all extraction costs and normal returns on produced capital have been accounted for. The value of a resource reserve is then set equal to the sum of the present value of expected future resource rent flows generated from extracting the resource over its reserve life.

This treatment is akin to recognizing that the value of the all capital employed in the mineral industries is not equal to the cost of investments. Normally, it is assumed

¹It will be calculated within the asset boundary of the System of National Accounts (SNA). According to the SNA (2008, §12.17), the asset boundary of a subsoil resource is limited to its proven reserve. The proven reserve of a subsoil resource is defined as its stock that is technically feasible and economically valuable for exploitation. The reserve data by commodity for Canada can be found in Statistics Canada CANSIM tables 153-0012 to 153-0028. Although different terminologies are used for describing reserves in these tables, such as “established reserve” for oil & gas and sulphur, “recoverable reserve” for coal and uranium, and “proven and probable reserve” for all others, they all refer to as the “developed reserves” defined as those that can be expected to be recovered through existing installations (wells or mines) under existing operating methods and economic conditions (Statistics Canada 2006, Text Box 3.1). As seen, the “proven reserve” and “developed reserve” are in principle the same concept.

that well-functioning markets will bring into equilibrium the cost of capital and its value—the present value of the stream of earnings that are produced by it. But on occasion this will not occur because of the scarcity of assets or imperfections in markets. When that occurs, capital in excess of that derived from the costs of investment is employed in the industry. And that is regarded as the case particularly in the resource sector where endowments cannot easily be changed by human activity—or at least not in the short run.

Two important parameters are required for valuing subsoil resources. One is the rate of return on produced capital that will be used for calculating the resource rent, and the other is the nominal discount rate that will be used for the net present value (NPV) of a resource reserve. SEEA (2012, p. 145) recommends that the rate of return on produced capital and the discount rate should be equal and suggests using an economy-wide interest rate derived from returns on government bonds as the rate of return that should be used on produced capital and the nominal discount rate as well. This is akin to choosing an arbitrary exogenous rate of return for estimating the value of produced capital services in the MFP estimation process—a practice that Statistics Canada does not follow in its productivity accounts for two reasons. The rate of return that is required is the rate that the capital markets would require to cover the cost of capital. Using a government bond rate involves understating the cost of business-sector capital since it involves greater risk. Secondly, its use generates estimates of surplus that are earned above requirements of capital markets that are difficult to interpret. They leave values of surplus across non resource industries that to be consistent with the approach adopted here should also be incorporated into the multifactor productivity accounts.

This paper uses an assumption that is in accord with the practice used in the Canadian Productivity Accounts (CPA). CPA calculates the internal rate of return on produced capital from the estimates of surplus and produced capital stock at an industry level. This paper assumes that, over a long term, produced capital earns on average the same rate of return in a mining industry and the non-mining business industries as a whole.² The internal rate of return on produced capital for the non-mining business industries as a whole can then be used in calculating the cost of capital services for produced capital in the mining industry. In turn, the resource rent in a mining industry can be calculated as the residual of the surplus estimated from the National Accounts minus the produced capital services used in this industry. This approach is consistent with that followed in the Productivity Accounts and profit remains zero for all industries except those using natural capital.³

²Baldwin and Gu (2007) show that this average derived from the Productivity Accounts closely approximates the cost of capital derived from the long-term corporate bond rate and the equity rate of return earned on Canadian equities.

³The SEEA (2012) suggests using an economy-wide interest rate derived from returns on government bonds as the rate of return on produced capital and the discount rate as well. The alternative of using an exogenously chosen rate of return that is suggested by the SEEA (2012) is not used here because of the difficulties in arbitrarily choosing an exogenous rate of return and

Once the resource rent as the surplus is estimated, an estimate of the value of natural capital that is the source of this surplus is derived from calculating the net present value of these surpluses. This is calculated using the estimates of resource reserves to estimate the years of remaining life at present extraction rates and then calculating the net present value of the surplus. The crucial parameter that is required for this analysis is the discount rate.

This paper adopts Hotelling's rule as the principle in the calculation of the NPV of subsoil resource reserves. Hotelling's Rule defines the optimal extraction path of non-renewable natural resources and predicts that the net price (unit resource rent) of a non-renewable natural resource is expected to increase at the rate of nominal interest that would be earned by an appropriate asset.⁴ Under Hotelling's Rule, the real discount rate becomes zero and the corresponding NPV of a subsoil resource reserve would reflect its value to a society if the source reserve is efficiently extracted.

Alternate choices of the discount rate have been suggested. For example, the SEEA (2012) assumes that the unit resource rent is expected to increase at the rate of general inflation. Under this assumption, the real discount rate used would equal the real rate of interest. In this case, the value of a resource reserve would be much smaller than that calculated using Hotelling's rule. This paper also provides an estimate of the value of natural capital coming from the Productivity Accounts that makes use of this assumption for the purposes of comparison.

The rest of the paper is organized as follows. Section 13.2 develops a framework for accounting for subsoil resources in production and wealth accumulation. Section 13.3 presents the empirical results for the Canadian mining industries and Sect. 13.4 concludes.

13.2 Framework for Accounting for Subsoil Resources

To isolate their contribution in production, subsoil resources are treated as a distinct factor of production in the same manner as labour and produced capital. Kendrick (1976) recommended that capital measures include machinery and equipment, structures, land, inventories and natural resource capital. Following the recommendation, a Hicksian neutral production function of subsoil resource extraction can be written as

$$Y = Af(L, Z^K, Z^N) \quad (13.1)$$

the inconsistencies that would result from not dealing with the surplus that would be generated in other industries.

⁴For a summary of the literature on whether resource prices have increased at this rate, see Miller and Upton (1985), Livernois (2009), and Kronenberg (2008).

where, the output (Y) is value-added based and a function of labour input (L), produced capital input (Z^K), and natural capital input (Z^N), augmented by productivity (A). For the production function to be well-defined, it is assumed that the marginal products of each factor are increasing ($\partial f/\partial L \geq 0$, $\partial f/\partial Z^K \geq 0$, $\partial f/\partial Z^N \geq 0$) at a decreasing rate ($\partial^2 f/\partial L^2 \leq 0$, $\partial^2 f/\partial (Z^K)^2 \leq 0$, $\partial^2 f/\partial (Z^N)^2 \leq 0$) and that all cross marginal products are increasing ($\partial^2 f/\partial L \partial Z^K \geq 0$, $\partial^2 f/\partial L \partial Z^N \geq 0$, $\partial^2 f/\partial Z^K \partial Z^N \geq 0$, $\partial^3 f/\partial L \partial Z^K \partial Z^N \geq 0$).

Equation (13.1) can be applied for the extraction of single or multiple subsoil resources. Logarithmically differentiating (13.1) yields

$$\frac{\dot{Y}}{Y} = \alpha_L \frac{\dot{L}}{L} + \alpha_K \frac{\dot{Z}^K}{Z^K} + \alpha_N \frac{\dot{Z}^N}{Z^N} + \dot{A} \quad (13.2)$$

where α_L , α_K and α_N denote the elasticities of output with respect to labour, produced capital and natural capital, respectively. These elasticities are not observable but can be derived by imposing the optimization conditions such that, for each factor of input, the value of its marginal products and its user costs are the same. Under the assumption of perfect competition and given output price (P^Y) and factor input prices (C^J), the output elasticities can be measured as

$$P^Y \frac{\partial Y}{\partial J} = C^J \Rightarrow \alpha_J \equiv \frac{\partial \ln(Y)}{\partial \ln(J)} = \frac{J}{Y} \frac{\partial Y}{\partial J} = \frac{C^J J}{P^Y Y} \equiv s_J, \quad \text{for } J = L, Z^K, Z^N. \quad (13.3)$$

Income and expenditure in extraction can be equated under the assumption of constant returns to scale, i.e.

$$P^Y Y = \sum_J C^J J = wH + cP^K K + \theta P^N N \quad (13.4)$$

where the labour cost is equal to the hours worked (H) multiplied by the nominal wage rate (w); the cost of produced capital is equal to its nominal stock value ($P^K K$) multiplied by the unit user cost of produced capital (c); and the user cost of natural capital is equal to its nominal stock value ($P^N N$) multiplied by the resource rent parameter (θ). Equations (13.3) and (13.4) show that the output elasticities in (13.2) can be replaced with the corresponding factor shares (s_L , s_K , and s_N) in the total value-added, i.e.

$$\frac{\dot{Y}}{Y} = s_L \frac{\dot{L}}{L} + s_K \frac{\dot{Z}^K}{Z^K} + s_N \frac{\dot{Z}^N}{Z^N} + \dot{A} \quad (13.5)$$

To use (13.5) for growth accounting, the growth of natural capital input and the resource rent associated with the use of natural capital need to be estimated.

13.2.1 Measuring Resource Rent

In this paper, the resource rent of natural capital is derived by using a residual value method.⁵ From (13.4) the resource rent (R) generated from extracting a subsoil resource is calculated residually as

$$R = \theta P^N N = P^Y Y - wH - cP^K K = OS - cP^K K \quad (13.6)$$

The data required for calculating the resource rent generated from single subsoil resource extraction include the corresponding gross operating surplus (OS) calculated as nominal value-added net of labour cost, nominal value of produced capital stock, and the unit user cost of produced capital.

The unit user cost of produced capital, which is equal to the sum of a rate of return on and a rate of depreciation of produced capital, needs to be exogenous to the mining industries in order to calculate the resource rent residually. There is no consensus in the literature on the choice of the exogenous rate of return on produced capital.⁶ One proposal that has been made is the borrowing costs. The borrowing costs in financial markets generally reflect the compensation to lenders for the provision of funds and the risk of loans not being returned. For example, a risk-free rate (the internal reference rate between banks) plus a risk premium of 1.5 % is used as the exogenous rate of return on produced capital in the Dutch national accounts for the calculation of resource rent in mining (Veldhuizen et al. 2012). Another example is the approach proposed in a cross country study by Brandt et al. (2013) for Organization for Economic Co-operation and Development (OECD), in which average extraction costs across countries are used to derive exogenously the resource rent of natural capital. Baldwin and Gu (2007) used a weighted average of the actual long-term debt costs and the equity rate of return earned in Canada for the purpose of examining how this approach compares to the endogenous estimate when deriving capital services and multifactor productivity growth in Canada. They find that the two are relatively similar for Canada.

There are several issues related to the use of exogenous rate of return on produced capital based on financial market information. First, using a flat exogenous rate of return will lead to high volatility in the measured resource rent and sometimes negative resource rent that may not accord with long-run expectations that are relevant for the derivation of the concept of the user cost of capital. Second, deriving a variable rate from financial market data that corresponds with longer run expectations is difficult because short run financial market fluctuations may not necessarily reflect long-run expectations. Third, a rate of return obtained from financial markets is usually an after-tax measure and needs to be converted into

⁵SEEA (2012) discusses various approaches for estimating resource rent and recommends the use of the residual value method.

⁶SEEA 2012 recommends that real long term government bond rates can be used if appropriate industry specific rates of return are not available, as is the case for many countries.

a before-tax measure; otherwise the resource rent would be overstated. Finally, it should be noted that for our purposes, consistency is required between the estimates of the mining sector and other industries. Industry revenues and costs may not be equal elsewhere when an exogenous rate of return is used and a “profit residual” may be generated across industries other than mining. While the “profit residual” is interpreted as the resource rent in a mining industry, it is more difficult to classify the reason or reasons for the residual elsewhere other than short run deviations from market clearing and therefore leads to unnecessary white noise in interpreting the estimates for users.

To overcome these issues, this paper uses an alternative way of splitting the operating surpluses into returns on produced capital and returns on natural capital (resource rent) than those suggested by the SEEA. Specifically, the internal rates of return on produced capital are adjusted such that produced capital in a mining industry earns the same rate of return as in the non-mining business sector on average over a long period.

13.2.1.1 Resource Rent at Commodity Level

The industry level at which MFP growth is estimated is more aggregated than the level of commodity data produced for the Environment Accounts at Statistics Canada and each mining industry at this level involves multiple resources that are estimated separately in the Environmental Accounts. While the latter involve more detailed data at the commodity level, they are not at the moment fully reconciled to the industry accounts that make up the basis for the multifactor productivity estimates. To calculate the resource rent at the industry level used in the productivity accounts, the gross operating surplus and the nominal value of produced capital stock at the commodity level are benchmarked to those at the industry level.

After the benchmarking, the internal rates of return on produced capital for the mining industries at the commodity level and the corresponding adjusted rates are then calculated. At the commodity level, data for produced capital at asset level and associated tax parameters are not readily available. Therefore, the internal rates of return on produced capital are calculated before tax and depreciation and have no asset details. Specifically, the gross internal rate of return on produced capital for commodity i and industry j is defined as

$$c_{ijt} = OS_{ijt} / (P_{ijt}^K K_{ijt}) \quad (13.7)$$

Resource rent at the commodity level in a mining industry is calculated as⁷

$$R_{ijt} = OS_{ijt} - \tilde{c}_{ijt} P_{ijt}^K K_{ijt}, \text{ with } \tilde{c}_{ijt} = c_{ijt} (\bar{c}_B / \bar{c}_{ij}) \quad (13.8)$$

⁷An alternative to (13.8) is to replacing c_{ijt} with r_{Bt} or its moving averages over a certain period.

where \bar{c}_B is the sample average of the gross internal rate of return on produced capital for the non-mining business sector, and \bar{c}_{ij} is that for the extraction of commodity i in industry j .

13.2.1.2 Resource Rent at Industry Level

At the industry level, more data is available; therefore the internal rate of return on produced capital after tax can be estimated. According to the user cost formula for produced capital developed in Christensen and Jorgenson (1969), the internal rate of return on produced capital in an industry (r_t) can be estimated as

$$r_t = \frac{OS_t + \sum_k (p_{kt-1} T_{kt} K_{kt} \pi_{kt} - p_{kt} T_{kt} K_{kt} \delta_k - p_{kt-1} K_{kt} \phi_t)}{\sum_k p_{kt-1} T_{kt} K_{kt}},$$

$$T_{kt} = \frac{1 - u_t z_{kt} - ITC_{kt}}{1 - u_t} \quad (13.9)$$

The asset-specific variables used in (13.9) include the user cost of produced capital (c_k), produce capital stock (K_k), asset price (p_K), depreciation rate (δ_k), capital gains (π_k), the present value of depreciation deductions for tax purposes on a dollar's investment (z_k), and the rate of the investment tax credit (ITC_k). Other variables are the effective rate of property taxes (ϕ) and the corporate income tax rate (u). We then use (13.9) to calculate the sample averages of the internal rate of return on produced capital for the non-mining business sector (B) and a mining industry (j) as

$$\bar{r}_B = \sum_{t=1}^n r_{Bt}/n, \quad \bar{r}_j = \sum_{t=1}^n r_{jt}/n$$

These sample averages can sensibly be related to expectations over the same period. It is usually expected that $\bar{r}_j > \bar{r}_B$ because \bar{r}_j includes returns on both produced and natural capital. If this is the case, it is assumed that produced capital earns the same rate of return on average over the sample period in these mining industries and in the non-mining business sector.⁸ The internal rates of return on produced

⁸Generally speaking, the expected rate of return on investment should be the same for different projects or across industries after adjusting for project- or industry-specific risks. Empirically, some investments, especially those in intangibles, are often not measured in the current capital stock measure. Also, project- or industry-specific risks are often different. As a result, the measured rates of return on capital stock over a long period are not necessarily the same across industries. However, the empirical evidence in Baldwin and Gu (2007) shows that the long-term average internal rate of return on capital in the Canadian total business sector is highly comparable with the long-term weighted average rates of interest on debt and equity in Canadian financial markets, which implies that the current coverage of capital stock in Canada may not be an issue in terms of the overall rate of return on capital. For a specific industry such as mining, its rate of return on capital may differ if the unmeasured investment and industry-specific-risks largely disproportionate

capital in the mining industries with $\bar{r}_j > \bar{r}_B$ are then adjusted by the ratio of the two sample averages. However, it can be the case that the actual data gives $\bar{r}_j \leq \bar{r}_B$ in the extraction of some subsoil resources. When this happens, the resource rent in these industries will be zero. For the industries with $\bar{r}_j > \bar{r}_B$, the adjustment is made as⁹

$$\tilde{r}_{jt} = r_{jt} \times \frac{\bar{r}_B}{\bar{r}_j} \text{ if } \bar{r}_j > \bar{r}_B \quad (13.10)$$

For a mining industry with $\bar{r}_j > \bar{r}_B$, the adjustment made by (13.10) does not change the pattern over time of the internal rate of return on produced capital (r_{jt}), but ensures that the sample averages of the adjusted rate of return on produced capital in the mining industry is the same as in the non-mining business sector, i.e.,

$$\text{Average } (\tilde{r}_{it}) = \sum_{t=1}^n \left(r_{it} \times \frac{\bar{r}_B}{\bar{r}_i} \right) / n = \frac{\bar{r}_B}{\bar{r}_i} \sum_{t=1}^n r_{it} / n = \bar{r}_B$$

In addition, the internal rates of return of produced capital derived from (13.10) are external to the mining industry of interest as it uses information of other industries. However, it uses national account information only.

The resource rents in a mining industry with $\bar{r}_j > \bar{r}_B$ is then residually calculated by subtracting the returns on produced capital calculated using the adjusted rates of return on produced capital, i.e.,

$$R_{jt} = OS_{jt} - \sum_k [T_{jkt} K_{jkt} (p_{kt-1} \tilde{r}_{jt} + p_{kt} \delta_{kt} - p_{kt-1} \pi_{jkt}) + p_{kt-1} K_{jkt} \phi_{jt}] \text{ if } \bar{r}_j > \bar{r}_B \quad (13.11)$$

13.2.1.3 Resource Rent Benchmarking

Due to data limitations at the commodity level, the resource rent estimate at the industry level is in general more reliable when the commodity-level resource rents are all positive. In this case the commodity-level resource rent is benchmarked using the industry-level resource rent as the control total, i.e.,

$$\tilde{R}_{ijt} = \frac{R_{ijt}}{\sum_i R_{ijt}} R_{jt}, \text{ for commodity } i \in \text{industry } j \quad (13.12)$$

from those for the total business sector. But if this were the case, the exogenous rate of return on capital used for a specific industry would also need to be modified to account for this.

⁹An alternative to (13.10) is to replacing r_{jt} with r_{Bt} or its moving averages over a certain period.

However, when the industry-level resource rent is zero or very small, it is recalculated as the sum of the resource rents at commodity-level,¹⁰ i.e.

$$R_{jt} = \sum_{i \in j} R_{ijt} \quad (13.13)$$

The resource rent generated from extracting a subsoil resource is taken here as the user cost or capital service of this natural capital asset. It is what the rental market for the assets would have to extract for the use of the natural capital if its use was rented out over the course of the year.¹¹

13.2.1.4 Resource Rent Decomposition

Let D be the physical extraction of a subsoil asset, and P^D be the unit user cost of the natural capital or the net price of the resource extracted at a point of time, we then have

$$R = P^D D \quad (13.14)$$

Exploration & Development expenditures have been capitalized as produced capital in the national accounts, implying that their returns have then been deducted in the calculation of the resource rent. As a result, the unit resource rent (P^D) reflects purely the value of a subsoil resource arising from its scarcity and quality of deposit.¹²

Similar to the user cost of produced capital, the resource rent can also be split into the depletion cost and returns on natural capital. Let P^N , δ^N and r^N denote the shadow price of, the depletion rate of, and the rate of returns on natural capital, respectively. The resource rent or the user cost of natural capital can then be

¹⁰In this situation, the bottom-up approach of (13.13) is superior to the top-down approach of (13.12). For example, assume in an industry with multiple resource extractions, the sample average of the internal rate of return on capital is low in the extraction of one resource but high in the extraction of all other resources. The industry-level internal rate of return can be low enough such that the resource rent derived directly using the industry-level data becomes zero simply because of the low internal rate of return in the extraction of one resource. In this case, a top-down approach will lead to zero resource rents for all resource extractions, while a bottom-up approach will result in zero resource rent for the resource extraction with low internal rate of return and positive resource rents for others with high internal rates of returns.

¹¹The user cost of using the reserve is the value of the surplus that is derived from its use. In the case of physical capital this involves both a depreciation of the asset and an opportunity cost of capital. These are both combined in the surplus actually derived from the natural capital asset.

¹²Let C be the total cost of a resource extraction including the cost of labour, produced capital and intermediate inputs, and P be the market price of the resource. The unit resource rent is equal to the market price net of the marginal cost of extraction ($P^D = P - \partial C / \partial D$) that is increasing in the degree of scarcity and the quality of deposit of the resource.

written as

$$P^D D = (\delta^N + r^N) P^N N = \underbrace{(P^N D)}_{\text{depletion cost}} + \underbrace{(P^D - P^N) D}_{\text{returns on natural capital}} \quad \text{with } \delta^N = D/N \quad (13.15)$$

13.2.2 Valuing Subsoil Resource Reserves

As there are often no readily available market prices for subsoil resource reserves,¹³ the Net Present Value (NPV) of the flow of natural resource rents is used here.¹⁴ The NPV method values a resource reserve from an ex-ante perspective. It converts the expected future streams of resource rents into the present value of a resource reserve. Let $E_t(d_{t+\tau})$ be the expected future nominal rate of return on a numeraire asset that is used for discounting future income flows, $E_t(\rho_{t+\tau})$ be the expected future growth rate of the unit resource rent, and T_t be the reserve life of a subsoil resource at a point of time, the NPV of the reserve of a subsoil resource becomes

$$NPV_{it} = P_{it}^N N_{it} = \sum_{\tau=1}^{T_{it}} \frac{E_t(P_{it+\tau}^D) D_{it+\tau}}{\prod_{s=1}^{\tau} (1 + E_t(d_{t+s}))} = \sum_{\tau=1}^{T_{it}} \frac{\prod_{s=1}^{\tau} (1 + E_t(\rho_{t+s})) P_{it}^D D_{it+\tau}}{\prod_{s=1}^{\tau} (1 + E_t(d_{t+s}))} \quad (13.16)$$

For notational simplicity we replace the period-specific discounts rates and growth rates of the unit resource rent in (13.16) with their annual averages over the reserve life, which yields

$$NPV_{it} = P_{it}^N N_{it} = P_{it}^D \sum_{\tau=1}^{T_{it}} D_{it+\tau} \left(\frac{1 + \rho_t}{1 + d_t} \right)^{\tau} \quad (13.17)$$

where

$$\rho_t = \text{Average}_{\tau=1}^{T_{it}} (E_t \rho_{t+\tau}), \quad \text{and} \quad d_t = \text{Average}_{\tau=1}^{T_{it}} (E_t d_{t+\tau}) \quad (13.18)$$

¹³Subsoil resources are traded both directly—in terms of transfers of land—and indirectly through the purchase of firms. While the value of the resources transferred is sometimes publically stated or calculated by the financial press, a large enough data base does not exist to allow use of these estimates here.

¹⁴The net present value (NPV) method is recommended in SEEA (2012) for the valuation of subsoil resource reserves.

Hotelling's rule¹⁵ suggests that the socially and economically optimal time path of a non-renewable resource extraction is one along which the resource price net of all extraction costs (unit resource rent) is expected to grow at the rate of return on investment (discount rate). That is

$$\rho_t = d_t \quad (13.19)$$

To understand the proposition, we assume that the representative agent chooses an extraction path to maximize the NPV of a resource reserve. The optimization can be written as

$$\begin{aligned} \max_{\{D_{it+\tau}\}_{\tau=1}^{T_{it}}} & \left(NPV_{it} = P_{it}^D \sum_{\tau=1}^{T_{it}} D_{it+\tau} \left(\frac{1 + \rho_t}{1 + d_t} \right)^{\tau} \right) \\ \text{s.t. } & \sum_{\tau=1}^{T_{it}} D_{it+\tau} = N_{it} \end{aligned} \quad (13.20)$$

The *Lagrangian* function for this problem can be written as

$$\Lambda = P_{it}^D \sum_{\tau=1}^{T_{it}} D_{it+\tau} \left(\frac{1 + \rho_t}{1 + d_t} \right)^{\tau} - \lambda \left(N_{it} - \sum_{\tau=1}^{T_{it}} D_{it+\tau} \right) \quad (13.21)$$

The first order conditions can be derived by taking derivative of (13.21) with respect to the physical extraction in each time, i.e.

$$\frac{\partial \Lambda}{\partial D_{it+\tau}} = P_{it}^D \left(\frac{1 + \rho_t}{1 + d_t} \right)^{\tau} - \lambda = 0, \quad \text{for } \tau = 1, \dots, T_{it} \quad (13.22)$$

It is required that $\rho_t = d_t$ for 13.22) to be hold. Otherwise, the current extraction is not optimal because the marginal profit of extraction and the marginal value of holding are not equal to each other. This is Hotelling's rule. Substituting $\rho_t = d_t$ into (13.17), (13.21), and (13.22) gives

$$\begin{aligned} \lambda &\equiv P_{it}^N = P_{it}^D \\ NPV_{it}^* &= P_{it}^N N_{it} = P_{it}^D \sum_{\tau=1}^{T_{it}} D_{it+\tau}^* = P_{it}^D N_{it} = R_{it} \widehat{T}_{it}, \quad \text{with } \widehat{T}_{it} \equiv \frac{N_{it}}{D_{it}} \end{aligned} \quad (13.23)$$

Therefore, along the optimal extraction path, the shadow price of a resource reserve is equal to the unit resource rent and both are expected to grow at the rate of nominal interest rate of a numeraire asset. The NPV of a resource reserve can then

¹⁵Hotelling's rule states the condition for the time path of a non-renewable resource extraction that maximizes the value of the resource stock. See Hotelling (1931) and Solow (1974) for details.

be calculated as the current resource rent multiplied by the number of periods of extraction at current level.

Hotelling's rule also implies that the rate of return on natural capital is zero. This can be seen using (13.15) when the shadow price of a resource reserve (P^N) is equal to the unit resource rent (P^D). So the benefits today (resource rents) fully reflect the cost of future loss (depletion costs).

In the above formulation, Hotelling's Rule was used to define the optimal extraction path of non-renewable natural resources to give the conceptual and theoretical framework for understanding and analyzing the depletion of non-renewable natural resources.

In support of the use to which the rule is being put here, Miller and Upton (1985) found that, for a sample of the U.S. oil & gas extraction companies, estimates of reserve values when calculated using Hotelling's rule account for a significant portion of their market values. They also compared the accuracy of using Hotelling's rule as opposed to two widely cited and public available alternatives—the the Securities and Exchange Commission (SEC) and Herald appraisals and reported that Hotelling's rule performed better in the valuation of the resource reserves values. This supports the use to which Hotelling's rule is being put here. It suggests that expectations are being formed to determine the values being estimated here using something approximating Hotelling's rule.

It is, however, the case that Livernois (2009) reports that empirical studies that examine the actual price trajectory find imperfect evidence that the actual trajectory of resource prices follows Hotelling's rule. But the question is not whether the trajectory follows Hotelling's rule exactly—but do the expected values using an approximation to this rule accord with values being created in markets, which is the criterion that accords with the spirit of measurement within the National and Productivity Accounts.

Kronenberg (2008) discussed factors that may lead to deviations of outcomes in the real world from those obtained from using Hotelling's rule. One category of these factors relates to the assumptions made for deriving the Hotelling's rule such as perfect competition, zero extraction cost, no technical progress, fixed stock of reserves, and constant market conditions. These assumptions can be relaxed. And in this paper, we do so by calculating the value of a resource by updating information continuously on the extraction cost, reserve stock, and market conditions, implying that the corresponding optimal extraction path of a resource reserve changes over time. The other category of these factors is institutional such as uncertain property rights and strategic interactions between suppliers and consumers. Although these institutional factors may lead to a market failure such that the actual extraction path is not socially optimal, valuing a resource reserve along its optimal path of extraction gives the value that can be achieved from the efficient extraction of a resource reserve.¹⁶

¹⁶Hotelling's rule is derived under the assumption of the existence of a representative agent. The assumption may not hold because some companies pay royalties to the owner of the resources and

13.2.3 Industry Level Measures

To this point, measures on the quantity and price for each natural capital asset have been derived. The industry level quantity and price measures are then aggregated from those for each asset using the Fisher formula. For the natural capital stock in a mining industry, its quantity and price indexes are calculated as

$$\begin{aligned} FQI_t^N &\equiv \frac{N_t}{N_{t-1}} = \sqrt{\frac{\sum_i P_{it-1}^N N_{it}}{\sum_i P_{it-1}^N N_{it-1}} \frac{\sum_i P_{it}^N N_{it}}{\sum_i P_{it-1}^N N_{it-1}}}, \\ FPI_t^N &\equiv \frac{P_t^N}{P_{t-1}^N} = \sqrt{\frac{\sum_i P_{it}^N N_{it-1}}{\sum_i P_{it-1}^N N_{it-1}} \frac{\sum_i P_{it}^N N_{it}}{\sum_i P_{it-1}^N N_{it}}} \end{aligned} \quad (13.24)$$

In the case of mining, the physical extractions are the service flows provided by the natural capital. The industry-level quantity and price indexes of natural capital service (input) can then be estimated as

$$\begin{aligned} FQI_t^{Z^N} &\equiv \frac{Z_t^N}{Z_{t-1}^N} = \sqrt{\frac{\sum_i P_{it-1}^D D_{it}}{\sum_i P_{it-1}^D D_{it-1}} \frac{\sum_i P_{it}^D D_{it}}{\sum_i P_{it-1}^D D_{it-1}}}, \\ FPI_t^{Z^N} &\equiv \frac{P_t^{Z^N}}{P_{t-1}^{Z^N}} = \sqrt{\frac{\sum_i P_{it}^D D_{it-1}}{\sum_i P_{it-1}^D D_{it-1}} \frac{\sum_i P_{it}^D D_{it}}{\sum_i P_{it-1}^D D_{it}}} \end{aligned} \quad (13.25)$$

The discrete approximation of the growth accounting formula can be derived from (13.2) as

$$\begin{aligned} \Delta \ln(Y_t) &= \bar{s}_t^L \Delta \ln(L_t) + \bar{s}_t^K \Delta \ln(Z_t^K) + \bar{s}_t^N \Delta \ln(Z_t^N) + \Delta \ln(MFP_t) \\ \text{with } \bar{s}_t^L &= (w_{t-1} L_{t-1} / Y_{t-1} + w_t L_t / Y_t) / 2, \quad \bar{s}_t^N = (R_{t-1} / Y_{t-1} + R_t / Y_t) / 2, \\ \bar{s}_t^K &= 1 - \bar{s}_t^L - \bar{s}_t^N \end{aligned} \quad (13.26)$$

Multifactor productivity (MFP) growth can then be estimated residually. It is noteworthy that the growth accounting framework (13.26) does not take into account the impact of changes in natural capital quality, so the derived MFP growth at this

some others do not. Given such heterogeneity among individual mining companies, the aggregate extraction path of a resource reserve may not be socially optimal even when each individual extraction path is optimal to each mining company. As a result, the Hotelling's rule may not hold exactly. But Miller and Upton's work suggests it holds approximately.

point only refers to the (natural capital) quality-unadjusted measure.¹⁷ Also, the impact of adding natural capital as an input into production on MFP growth relies on the relative growth of produced and natural capital. It raises MFP growth when the natural capital growth is lower than that for produced capital and vice versa.

13.3 Empirical Results for Canadian Oil and Gas Extraction

In this section, the growth accounting framework developed in the previous section is applied for the Canadian oil and gas mining industry as an experimental analysis. The commodity (asset) level data on the gross operating surplus and nominal produced capital stock for the mineral sector is compiled by the Environment Accounts and Statistics Division of Statistics Canada based on various data sources.¹⁸ These data are benchmarked to the industry level data first and then the benchmarked data are used for the calculation of the resource rents at the commodity level. The quantity measures of the stock, depletion and addition of each subsoil resource reserve are obtained from CANSIM tables 153-0012 to 153-0015. Combined with the estimates of resource rents, these data are used for the calculation of reserve value at the commodity level and the quantity and price indexes of natural capital stock and natural capital input at the industry level. The industry level data of value-added, labour compensation, labour and produced capital inputs come from the KLEMS database used in the Canadian Productivity Accounts (CPA) and the industry level geometric-based nominal produced capital stock data come from CANSIM table 031-0002.¹⁹ The gross operating surplus and the nominal capital stock data at both industry and commodity levels are used for estimating the resource rents at both commodity and industry levels. A zero real discount rate is used throughout our experimental assessment. Given that the natural capital input is measured by the amount of physical extraction, the choice of the discount rate has no impact on the measurement of MFP growth. However, the measured value of natural capital stock is much larger under the Hotelling's rule (zero real discount rate) than

¹⁷Firms need to dig deeper and/or extract more waste to extract the same amount of mineral or energy content due to declining in natural resource quality. As a result, technical progress would be understated by the quality-unadjusted MFP growth. As the quality adjustment may involve some major data development, we will address this issue in a separate paper.

¹⁸The capital stock data by both industry and commodity does not include land and inventories due to lack of measures or sufficient quality.

¹⁹The business sector is defined differently in CPA and in CANSIM table 031-0002. In CPA, the business sector combines the business establishments of the North American Industry Classification System (NAICS) codes 11–81, while in CANSIM table 031-0002 it covers the all industries less public administration (NAICS 91), education (NAICS 61) and health (NAICS 62). To ensure the two sets of business sector data are consistent, this paper subtracts education (NAICS 61) and health (NAICS 62) from the CPA business sector data. After the adjustment, the coverage difference between the two definitions is minimal.

those with the discount rate being 4 %.²⁰ This discount rate is currently used in the Statistics Canada Environmental Accounts and also in many other national statistical agencies.

Oil & gas extraction involves the extraction of natural gas, crude oil and crude bitumen. Natural gas liquids are included in the asset category of natural gas.²¹ The estimates on the volume of reserve and extraction for each type of resources are presented first. The estimates of the nominal value of reserve and the resource rent of the extraction are presented next. The volume estimates of reserves are then aggregated across different types of resources to derive total natural capital stock, while the extractions are aggregated to derive the flow of services for the natural capital (or natural capital input), using weights based on resource rents. Finally, the contribution of the natural capital to output and its effect on MFP estimates are presented.

13.3.1 Resource Reserve and Extraction

The established reserve of oil & gas in Canada has experienced a large compositional shift towards crude bitumen. As shown in Fig. 13.1, over the period of 1981–2009, the established reserve trended down slightly for both natural gas and

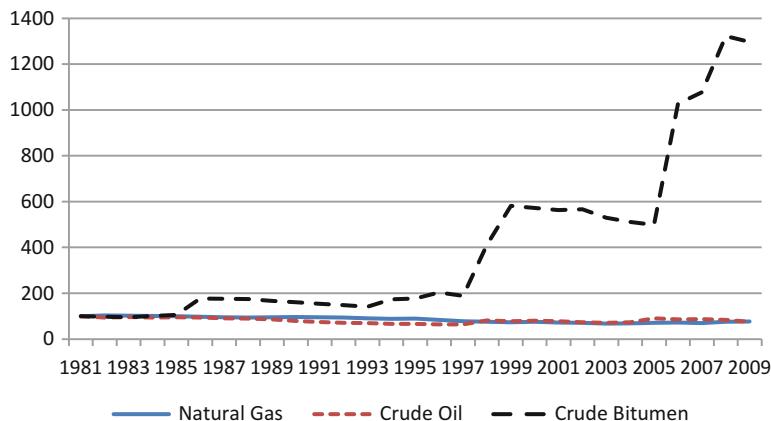


Fig. 13.1 Trend in established reserve, oil & gas, 1981 = 100. Source: Statistics Canada, authors' calculation based on CANSIM tables 153-0012, 153-0013, 153-0014, and 153-0015

²⁰See Appendix Table A.1 for Oil & Gas.

²¹The volume of natural gas liquids is approximately 1/600 of the gaseous volume at atmospheric conditions. We apply this conversion factor to make natural gas liquids and normal natural gas additive in volume.

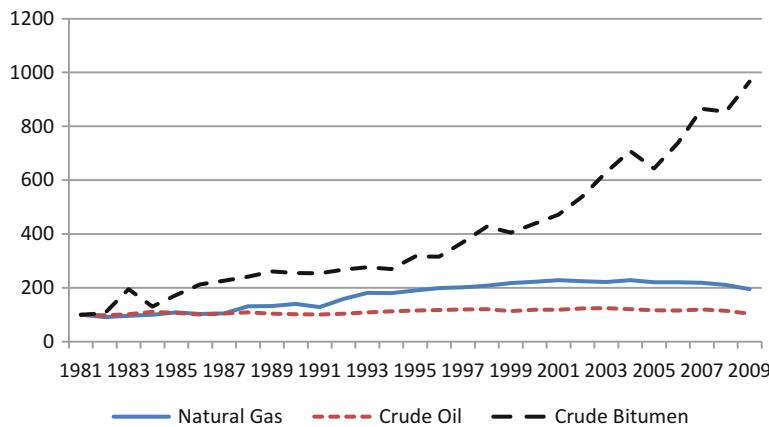


Fig. 13.2 Trend in extraction of oil & gas reserves, 1981 = 100. Source: Statistics Canada, authors' calculation based on CANSIM tables 153-0012, 153-0013, 153-0014, and 153-0015

crude oil. It dropped by about 25 % for both natural gas and crude oil, respectively, over the whole sample period. At the same time, the established reserve of crude bitumen increased dramatically, especially in the periods from 1997 to 1999 and after 2005. It increased by more than 12 times, or about 9.6 % per year on average.

Unlike the pattern over time of the established reserve, the extraction of all three oil & gas resources has increased, although at quite different paces (Fig. 13.2). Over the 1981–2009 periods, extraction grew by about 2.4 % per year for natural gas, 0.1 % per year for crude oil, and 8.4 % per year for crude bitumen.

13.3.2 Resource Rent and Reserve Value

Resource rent for oil and gas extraction is calculated directly using the industry-level data, and the resource rents at commodity level are benchmarked to the industry-level estimate of resource rent. Figure 13.3 presents the estimated value of oil & gas reserves and the resource rent from the extraction of oil & gas over the period of 1981–2009. As shown, the patterns over time of the reserve value and the resource rent are quite close to each other. Both stayed low and stagnant before 1999 and then grew rapidly thereafter. The annual resource rent declined by 2.5 % per year over the 1981–1999 periods and increased by 17.7 % per year over the 1999–2009 periods. The corresponding growth rates for the reserve value were –4.2 and 22.9 % per year for the two periods, respectively.

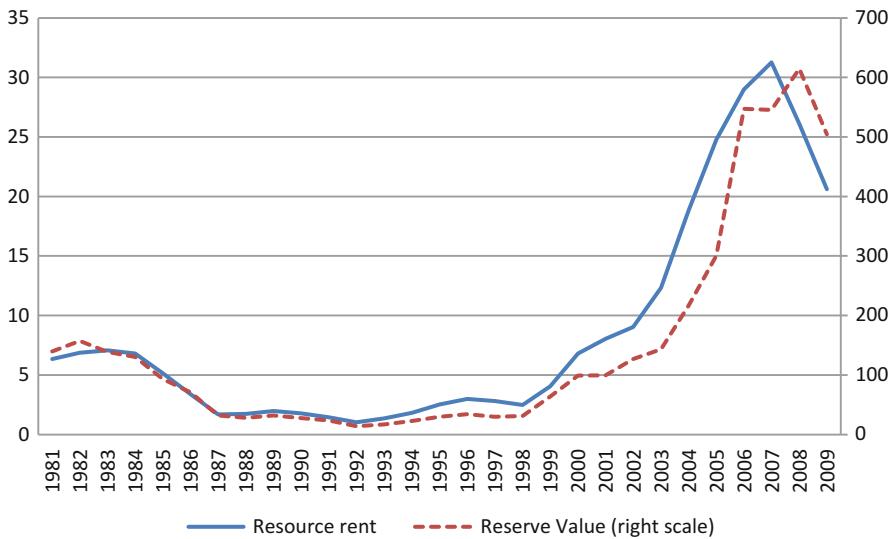


Fig. 13.3 Resource rent and reserve value, oil & gas, in billions of current dollar. *Source:* Statistics Canada, authors' calculation based on KLEMS database and environment accounts

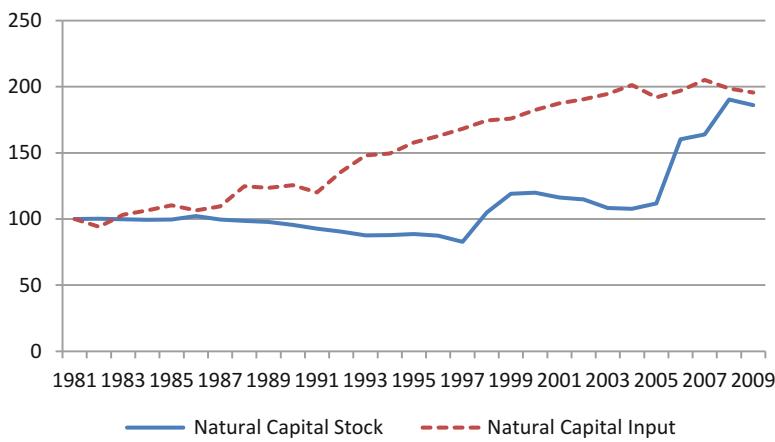


Fig. 13.4 Trend in volume of natural capital stock and natural capital input, oil & gas, 1981 = 100. *Source:* Statistics Canada, authors' calculation based on KLEMS database and environment accounts

13.3.3 *Natural Capital Stock and Natural Capital Input*

The natural capital input in this industry trended up steadily without major interruptions (Fig. 13.4). It grew by 2.4 % per year on average over the period of 1981–2009. At the same time, the pattern over time of the natural capital stock

is quite different from that of the natural capital input. The natural capital stock trended down gradually and dropped by about 17 % before 1997, reflecting the down-trending movements in natural gas and crude oil reserves. After 1997, the natural capital stock had a pattern over time similar to that of crude bitumen. It increased largely during 1997–1999 and after 2005, and decreased moderately during the period of 2000–2005.

13.3.4 MFP Growth

In our growth accounting framework, adding natural capital has no impact on output (value-added) growth and the contribution of labour input. However, the income share and hence the contribution of produced capital input will be reduced; as a result, MFP growth would be impacted if the produced capital input and the natural capital input grow at different paces.

As shown in Fig. 13.5, MFP growth in oil & gas extraction was positive before 1993 and became largely negative after 1993. Note that the impact of adding natural capital in the growth accounting framework on MFP growth is small before 1993 and large thereafter. Specifically, adjusting for natural capital changed the annual MFP growth from 1.8 to 2.0 % before 1993, and from -5.1 to -4.0 % after 1993. Overall, by including subsoil resources, MFP declines by 1.5 % per year over the 1981-to-2009 period, compared to a 2.2 % decline without including these resources.

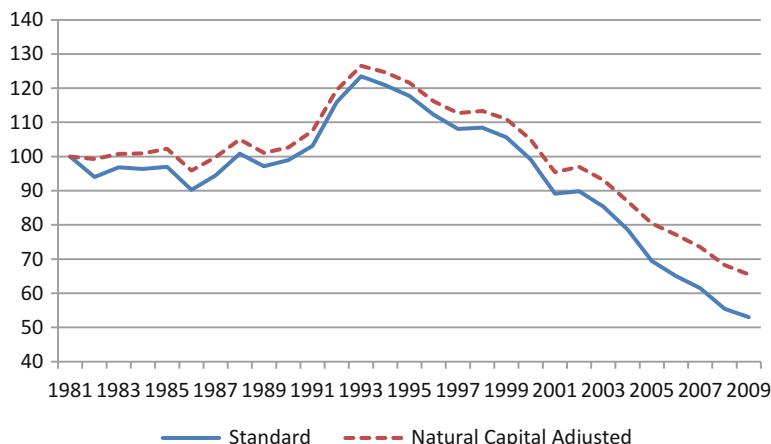


Fig. 13.5 Alternative measures of multifactor productivity, oil & gas, 1981 = 100. *Source:* Statistics Canada, authors' calculation based on KLEMS database and environment accounts

Table 13.1 Source of value-added growth, oil and gas extraction

	1981–2000	2000–2008	1981–2009
Value-added growth (log), annual average, %	3.22	0.39	2.31
Contribution, percentage points			
Labour input	0.08	0.84	0.32
Produced capital input	2.45	4.64	3.16
Natural capital input	0.43	0.16	0.34
MFP	0.26	-5.25	-1.51
MFP growth (log), annual average before adding natural capital, %	-0.04	-6.96	-2.27

Source: Statistics Canada, authors' calculation based on KLEMS database and environment accounts

13.3.5 Natural Capital Contribution to Value-Added Growth

The contribution of the natural capital input to the industry value-added growth is moderate in oil & gas extraction. Over 1981–2009, the log growth of value-added in oil & gas extraction was about 2.3 % per year, of which about 0.3 percentage points per year or 15 % came from the growth in the natural capital input (Table 13.1).

13.4 Conclusion

To recognize subsoil energy and mineral resources as a capital input into the production process, this paper presents a growth accounting framework that allows the derivation of measures on natural capital stock and natural capital input in the mining industries and provides a better understanding of contribution of natural capital to economic growth and the impact of adding natural capital on productivity measurement.

The empirical results suggest a significant contribution of natural capital to the real value-added economic growth in the Canadian oil and gas extraction. However, the impact of adding natural capital in the growth accounting on the measured MFP growth changes over time. It is small before 1993 and becomes large thereafter.

Acknowledgements We would like to thank John Baldwin, Wulong Gu, Michael Wright of Statistics Canada, Pierre-Alain Pionnier of OECD, Michael Smedes of Australian Bureau of Statistics, Vernon Topp of Australian Productivity Commission, Erik Veldhuizen of Statistics Netherlands, and Carl Obst of the London Group for their valuable comments and suggestions. Also thank participants of 2013 CANSEE (Canadian Society for Ecological Economics) conference at York University, Toronto, and 2014 NAPW (North American Productivity Workshop) VIII Conference at Ottawa/Gatineau, for helpful discussions. All errors are our own.

A.1 Appendix

	Value at 0 % discount/Value at 4 % discount
Total	1.49
Natural gas	1.38
Crude oil	1.21
Crude bitumen	1.80

Source: Statistics Canada, authors' calculation based on KLEMS database and environment accounts

	1981–2000	2000–2008	1981–2009
<i>Annual average cost share, %</i>			
Labour	12.8	9.6	11.8
Produced capital	70.8	64.9	68.5
Natural capital	16.5	25.5	19.7
<i>Average annual input growth (log), %</i>			
Labour	1.87	9.17	4.22
Produced capital	3.57	7.17	4.73
Natural capital	3.16	0.78	2.40

Source: Statistics Canada, authors' calculation based on KLEMS database and environment accounts

References

- Baldwin J, Gu W (2007) Multifactor productivity in Canada: an evaluation of alternative methods of estimating capital services. *The Canadian Productivity Review*, no. 009, Catalogue no. 15-206-XIE, Statistics Canada
- Brandt N, Schreyer P, Zipperer V et al (2013) Productivity measurement with natural capital. OECD Economic Department Working Paper, no. 1092
- Christensen LR, Jorgenson DW (1969) The measurement of U.S. real capital input, 1929–1967. *Rev Income Wealth* 15:293–320
- Hotelling H (1931) The economics of exhaustible resources. *J Polit Econ* 39(2):131–75
- Kendrick JW (1976) The formulation and stocks of total capital. NBER, New York
- Kronenberg T (2008) Should we worry about the failure of the Hotelling rule. *J Econ Surv* 22(4):774–93
- Livernois J (2009) On the empirical significance of the Hotelling rule. *Rev Environ Econ Policy* 3(1):22–41
- Miller MH, Upton CW (1985) A test of the Hotelling valuation principle. *J Polit Econ* 93(1):1–25
- SEEA (2012) System of Environment-Economic Accounting: central framework. United Nations, New York
- SNA (2008) System of National Accounts. United Nations, New York

- Solow RM (1974) The economics of resources or the resources of economics. *Am Econ Rev* 64(2):1–14
- Statistics Canada (2006) Concepts, sources and methods of the Canadian system of environmental and resource accounts. Catalogue no. 16-505-GIE
- Veldhuizen E, de Haan M, Tamrizeen M, van Rooijen-Hoesten M et al (2012) The Dutch growth accounts: measuring productivity with non-zero profits. 32nd General Conference of the International Association for Research in Income and Wealth, Boston

Chapter 14

Balancing Incentives: The Development and Application of a Regulatory Benchmarking Model

Roar Amundsveen and Hilde Marit Kvile

Abstract The main contribution in this paper is a presentation and discussion of issues that arise in the practical application of a regulatory benchmarking model. We describe the regulatory benchmarking model for electricity distribution companies in Norway, and we focus on how different choices influence different incentives for the companies. These choices cover methodology, modelling assumptions and variables, but also how the benchmarking results are applied in the regulatory model. The benchmarking model is only one part of the regulatory model for setting revenue caps. This discussion shows some of the trade-offs that have to be considered in this process, and sheds some light on why regulators may deviate from optimal text-book solutions.

Keywords Regulation • Benchmarking • Revenue caps • Incentives • Electricity distribution

14.1 Introduction

Norway has 144 distribution system operators (DSOs) serving a population of about 5.1 million people. The DSOs are institutional monopolies that are regulated by the Norwegian Water Resources and Energy Directorate (NVE), which is a directorate under the Ministry of Petroleum and Energy. One of NVE's mandates is to promote efficient energy markets and cost-effective energy systems, and since 1997 revenue cap regulation has been one of the means to achieve this. As many other regulators, NVE has applied benchmarking methods to evaluate the companies' performances as part of the revenue calculation. Data Envelopment Analysis (DEA) has been used since 1998. However, the regulatory models have been revised and improved a number of times.

R. Amundsveen (✉) • H.M. Kvile
The Norwegian Water Resources and Energy Directorate, Oslo, Norway
e-mail: roam@nve.no; hkv@nve.no

Based on the results from regulatory case studies in Europe, Weyman-Jones (2006) concluded that the regulatory application of efficiency and productivity analysis in practice has not kept up with the development of the theoretical and empirical methodology. The author's impression is that sample size, variable choice, model specification and choice of methodology have been governed by different objectives from those of the theoretical literature. While the theoretical research interest is in performance measurement for improved efficiency, the regulatory purpose has been feasible capture of economic rent.

In a review of the regulatory reforms in the UK, Waddams Price (2000) concluded that the effects of the high powered incentives were limited by distributional issues. On the political agenda, a question was how to balance between efficiency incentives themselves and the equity concerns of forcing the companies to share their effects with its customers. This was based on allegations that customers had not received a sufficient share of the benefits.

Haney and Pollit (2012) recognize that a regulatory application of benchmarking is an interactive process which involves negotiation and is subject to external ex post scrutiny. As such it would be wrong to condemn simple benchmarking models as "wrong" or "ineffective" per se. They may be useful negotiation devices for the regulator to discover best practice. If the results from these models bear little relation to reality then there is room for challenge within the process.

In general, the literature of regulation contains detailed descriptions of models for optimal prices, different regulatory schemes from cost-of-service to high-powered incentive schemes, benchmarking models based on different assumptions etc. However, there are less descriptions of how regulators actually develop regulatory models, for example judgments regarding choice of inputs and outputs, methodology etc. This may be too detailed to be part of the literature, but may also be regarded as internal regulatory matters that should not become public.

The Norwegian regulatory model is transparent in the sense that all data, assumptions and calculations are published, and it is possible to check all results in retrospect. This model is therefore an ideal case study in order to highlight the importance of the regulator being aware that every element in the model will affect the companies' incentives, and how it affects the incentives. By presenting the benchmarking model for the Norwegian DSOs we will discuss this process of balancing of incentives, and we will give examples of how and why regulatory choices may deviate from the ideal text-book solutions.

This paper is organized as follows, in Sect. 14.2 we discuss some of the core elements of regulation. In Sect. 14.3 we describe the outline of the Norwegian regulatory model. In Sect. 14.4 we describe the calculation of the cost norm where the benchmarking model is applied. During the descriptions in Sects. 14.3 and 14.4 we will discuss how the different decisions affect the incentives of the network company and which interests are balanced against each other. In Sect. 14.5 we summarize the discussion.

14.2 The Core Elements of Regulation: A Norwegian Experience

Implementation and development of a sustainable regulatory regime depends on several factors. Based on the Norwegian experience, the most important are:

- An appropriate legal framework
- Clear and defined objectives
- An efficient regulatory model framework
- Compliance monitoring
- The application of sound regulatory “craftsmanship” over time

In Norway the legal framework of the regulation of the electricity distribution companies is based on the Energy Act from 1990.¹ The main objective of the Energy Act is (among other factors) to ensure that transmission and distribution of electricity occur in a socially rational manner. This is the goal for the regulator, and illustrates the point that regulators are not interested in efficiency per se, but in using benchmarking as a tool to reach the goal set in the Energy Act. A range of regulations² based on the energy act limits the operations of the companies.

Regulatory objectives are set in one of the regulations following the Energy Act.³ It says that NVE shall “*... calculate yearly revenue caps, and revenues shall over time cover the company’s costs operation and depreciation, and give a reasonable rate of return on the capital given efficient maintenance, development and utilization of the grid*”.

NVE develops the regulatory framework given the objectives of acts and regulations. The regulatory model will be described in Sect. 14.3. This describes the economic regulations that are designed to give proper incentives for the companies. But incentives can only influence some of the behaviour of a company, so there is need for direct regulations and compliance monitoring to ensure that the companies comply with all direct regulations. NVE also has the responsibility of monitoring.

Even though much is given in the acts and regulation, the regulatory practice over time is also an important factor. It is important that the industry regards the regulator as reliable, transparent, predictable and non-discriminating. Fast and consistent case proceeding over time is one factor to achieve this.

¹<http://lovdata.no/lov/1990-06-29-50>

²<http://lovdata.no/referanse/hjemmel?dokID=NL/lov/1990-06-29-50>

³<http://lovdata.no/dokument/SF/forskrift/1990-12-07-959> §4-4 b)

14.3 The Regulatory Model for Norwegian DSOs

14.3.1 Allowed Revenue

One of the legal regulations based in the Energy Act regulates the reporting of economical and technical data, revenue caps and tariffs. This regulation is the legal basis for how NVE calculates allowed revenue for each company yearly, following this formula:

$$AR_i = RC_i + PT_i + TL_i - VOLL_i$$

AR_i is the allowed revenue for DSO i and RC_i is the revenue cap for DSO_i. PT_i are pass through costs; they are considered to be outside of company control. They are therefore neither benchmarked nor included in the revenue cap model by NVE.⁴ It is important for the regulator to acknowledge that all costs are not controllable by the companies. TL_i is a mechanism for removing time lag for investments. $VOLL_i$ is value of lost load. This mechanism was introduced in 2002 to calculate the socio-economic cost for interruptions. For every interruption the $VOLL$ is calculated by cost functions, and the value depends on the type of customers that is affected, duration and time of the interruption. Deducting $VOLL$ from allowed revenue makes sure companies also consider quality of supply against cost efficiency.

The mechanism for removing time lag on investments (TL) is included because the costs used in calculating the revenue cap are 2 years old (e.g. 2011-data for 2013-revenue caps). These are the most recent data NVE has when calculating ex-ante revenue caps. Using 2-year-old capital cost in the revenue cap may give disincentives for investments. This has been a general challenge in regulation, and over the years, regulators have applied different solutions to increase incentives for investment. One example is an adjustment for investments that is calculated based on previous investments. From 2009, NVE removed the time lag on capital entirely by introducing this TL mechanism. As an example, when allowed revenue for 2013 was calculated (in December 2014), the actual depreciations and book values for 2013 were known. The difference between these capital costs and the values for 2011 (that were used calculating revenue caps for 2013) were added to the allowed revenue. This means that companies can include estimated capital costs from investments in the tariff base already from the year of the investment. It also means that for the first 2 years of an investment, the capital is not benchmarked and the companies know they will receive the regulatory rate of return.

When the allowed revenue is calculated, this is compared to the actual revenue. The latter is the total revenues from tariffs. In the model, the companies will not lose revenue if actual revenue is less than allowed revenue for a given year. NVE

⁴These are property taxes and costs incurred in higher network levels. In addition the companies are allowed to cover some of their R&D costs directly in the allowed revenue. In this way NVE gives incentives for R&D among companies.

calculates a surplus or deficit balance for each company yearly. Surplus plus interest rates must be paid back to the consumer through lower tariffs; deficit may be collected from consumer (optional with interest rates) through higher tariffs. The company must target this balance towards zero over time, but the mechanism gives the companies some degrees of freedom when setting tariffs and it avoids tariffs varying too much from year to year. It also illustrates that the regulator has a long time horizon in its regulation, and makes the regulation model seem predictable for the companies.

A large part of the incentives for cost efficiency comes from the revenue cap model, but we see that both incentives for investments and quality are included when allowed revenue is calculated as well.

14.3.2 Revenue Caps

The revenue cap is the main element entering the calculation of allowed revenue, and many incentives are balanced in the making of this model. The revenue cap formula is described in the same legal act as allowed revenue above.

$$RC_i = (1 - \rho) C_i + \rho C_i^*$$

RC_i is the revenue cap for DSO_i, ρ is a scalar with possible values between 0 and 1, for the time being set to 0.6. C_i is the cost base for DSO_i and C_i^* is the cost norm for DSO_i. The cost base is calculated from the company's own costs and the cost norm is calculated based on other companies' costs. The calculation of the cost norm will be described in Sect. 14.4.

The following cost elements are included in the cost base:

- Operation and maintenance costs
- Energy losses, calculated as the company's reported volume of losses calculated with an area price per MWh.
- Value of lost load, VOLL. VOLL is included in the cost base, even though it is technically a reduced revenue, not an actual cost.
- Capital costs are calculated as depreciations and calculated return on the regulatory asset base (RAB) by 31.12. The RAB is based on book values from the company's accounts. The return is a regulatory rate of return calculated from a WACC model. A further description of the WACC model is found in Langset and Syvertsen (2013).

All costs are updated every year, so the companies know what costs enter the revenue calculations.

Using both cost base and cost norm implies a sharing of risk and profit between the company and their customers, and the size of ρ decide the strength of the cost efficiency incentives (Shleifer 1985). When $\rho = 0$, we have a cost plus model with no incentives for cost efficiency. $\rho = 1$ is a pure yardstick model where the

companies' revenues are completely independent of their own costs, giving very strong incentives for cost efficiency. Setting a reasonable value for ρ will depend on for instance quality of data and trust of the model. The ρ in the Norwegian model is also a variable set in the legal regulation and it is 0.6 in order to let incentives for cost efficiency be a bit stronger than the confidence that you will recover 40 % of your own cost through the cost base. In April 2014 a group of experts presented a report on the structure of the electricity distribution company industry ordered by the Ministry of petroleum and energy (Ministry of Petroleum and Energy 2014). They recommended increasing ρ to 0.7 to strengthen the incentives for cost efficiency. On the other hand, such an adjustment implies less security for companies, and companies with low scores from the benchmarking can get cash flow problems. These interests must be balanced against each other.

14.4 The Cost Norm

This section includes a description of the benchmarking model and how the benchmarking results are applied for calculating the revenue caps. Some incentives in the regulatory model are based in the elements of the model described in Sect. 14.3, but most of the incentives are based in the application of the cost norm. The calculation of the cost norm is not described in the legal regulation; that merely states that the cost norm shall be calculated using comparative analysis that must take into account relevant differences in the operating conditions. Still, when NVE changes elements in the cost norm calculation, we treat it similarly to changes in the regulations. We ensure a proper communication with the industry and send the suggestions on public hearing as part of our definition of sound regulatory craftsmanship.

The cost norm is calculated in three stages. Choices in one stage can create challenges in other stages and weaknesses in one stage can be reduced in other stages.

1. Benchmarking model based on DEA
2. Adjustment of DEA scores due to heterogeneity in operational environments
3. Calculation and calibration of cost norms

For NVE, DEA has been the main methodology in the calculation of revenue caps since 1998, but other methodologies such as COLS and SFA have also been used in analyses and in model development. Kittelsen (1994) recommended NVE to use DEA because there was limited information about the properties of the production function. In the beginning, the industry regarded the benchmarking model as "a black box", but this view has gradually vanished during the years. Today the industry seems to be quite confident in the use of DEA. However, the industry supports the search for better benchmarking models.

From 2007, NVE included variables to capture operational environments due to heterogeneity in climate, topography and geography (also called Z-variables) as

outputs in the DEA model. This tended to put too much emphasis on some of the geographical variables; some companies became peers in geographical dimensions, and for some companies geography decided the whole DEA result. Therefore, NVE changed to a second stage procedure in 2010 (NVE 2009). In theory, it may be better to analyse all the relevant variables in **one** single optimisation. Alternative benchmarking methods that can consider Z-variables in one stage and are used by regulators in other countries are for example StoNED method (Kuosmanen 2012) applied by the Finnish regulator. NVE has chosen to keep the DEA model and has developed a new and improved approach of the second stage. This is described in Sect. 14.4.2. One large advantage of using DEA was that the industry knew this method well, and NVE considered that with the application of a second stage DEA approach it was possible to achieve the regulatory goals.

14.4.1 Stage 1: The DEA Model

NVE's approach when choosing variables for a benchmarking model is that the variables must be conceptual, intuitive, significant and feasible. NVE applies a DEA model with one input and three outputs.

14.4.1.1 Input

It is possible to define inputs as quantities or as monetary values. In the table below we present different possible inputs in a model for electricity distribution companies (Table 14.1):

Inputs as quantities are suitable if factor prices are unknown, or vary between the different companies. Quantities may also be favourable in international studies. They are easier to compare between countries than monetary values, due to different accounting standards, different price levels etc.

Table 14.1 Input

Input (quantity)	Factor price	Monetary values
Man-years	Price per man-years	Total wages
Goods and services	Unit price	Goods and services
Network losses MWh	Price per MWh	Network losses
Interruptions MWh	Price per MWh	Socio-economic cost of interruptions
Capital (book values)	Depreciation rate and rate of return	Capital cost
Capital (replacement values)	Depreciation rate and rate of return	Capital cost
Capital (network length)	Price per km	Value of network capital
Capital (transformer capacity MVA)	Price per MVA	Value of transformer capital

The advantage of monetary outputs is that it is possible to add different inputs together, given that the factor prices are approximately the same between the companies. This is an advantage because a company for instance can choose to use its own employees, which would be reflected in the number of man-years (or working hours) and total wages. Alternatively, the company can choose to buy this service in the market. These two alternatives, or the combination of them, can contribute to the same level of output. The ideal combination can depend on the local market situation.

In the Norwegian model we have chosen monetary inputs, and we have chosen to add all cost into one input. The total cost is similar to the cost base⁵ and is the sum of:

- operation and maintenance cost (OM),
- cost of energy losses,
- value of lost load (VOLL),
- depreciations and
- regulatory rate of return⁶ on regulatory asset base (RAB)

NVE has chosen to add these into one input to give the companies incentives to view all costs in context when making decisions. Companies face trade-offs between the different costs. They can choose to increase maintenance and thereby delay a capital investment. They can choose to build more underground cable to reduce interruptions in the supply. It is also possible to invest to reduce network losses or choose a less expensive investment that will increase the network losses. Each company has a range of available input combinations in order to produce grid services as efficient as possible. If the regulator chooses to treat the different cost elements differently, it may give incentives to favour certain costs. E.g. if capital is not benchmarked it gives incentives to reinvest too early. NVE thinks that applying all costs in one input gives the companies incentives to lower their total costs in the best way.

Calculating capital costs is in our view one of the hardest issues, and how they are calculated has a strong effect on incentives for investments. In the Norwegian model, the regulatory asset base (RAB) is the book values from the companies' accounts. The advantage of using book values is that it reflects what the companies actually have paid for their assets, and all companies have to follow the same accounting rules. This increases comparability. However, using book values means that the input is affected by the age of the assets. Two companies can have exactly the same assets, costs of loss, OM and VOLL, but if one company has older assets than the other, the capital costs, and therefore the total cost, will be lower. The DEA results will therefore reflect both inefficiency and an age effect.

⁵In the input of the DEA we also include capital that have been financed through contributions. This is because these outputs are included in the companies' data, therefore we need to add the costs as well.

⁶We use the same WACC model as in the cost base.

It has been a critique to the regulation model since 2007 that the model gives disincentives to invest in a time where the need for investments is large.⁷ NVE has maintained the use of book values in the cost base by arguing that in the long run, all companies must invest and the age effect will disappear. However, NVE has recognized the challenge companies face when much of the cash flow comes towards the end of an asset's lifetime. During 2012, NVE considered changing the measure for capital in the DEA model to a replacement cost that would be independent of age of the assets. NVE wanted to keep the book values in the cost base, though, but combining book values in the cost base with an age independent measure in the benchmarking gave too strong incentives for reinvestment. Every reinvestment would be profitable, even if they were unnecessary since reinvestments would increase the cost base but not influence the cost norm. Even if the age-independent capital produced a more "correct" measure for efficiency, it was not applicable in the regulatory framework. Instead NVE changed elements in the third stage of the calculation of the cost norm to reduce some of the age effect in stage one. This illustrates how changing one detail in one end can change incentives in the other end. We will get back to this issue when describing stage 3 of the cost norm model.

14.4.1.2 Output

The table below shows different measures for the outputs of an electricity grid company, some are exogenous and others are endogenous to the company (Table 14.2):

In theory, we prefer exogenous outputs. We use an input minimizing DEA model where the companies should minimize their input given the output. Then it is unfortunate if the companies are able to influence the level of the outputs. However, even if length of network and number of transformers may seem endogenous, other direct regulations imply that they in reality are not. The DSOs have a duty to connect all customers and producers that demand it, and investments of new grid are basically driven by the external factors of supply and demand.

Table 14.2 Output

Exogenous to the companies	Endogenous to the companies
Demand for power (MW)	Length of high voltage network
Demand for connection (customers)	Length of low voltage network
Demand for energy (MWh)	Transformers and substations
Transport distance	Transformer capacity
Geography/Z-factors	Age of assets

⁷NVE has assumed the need for investments (for distribution companies) to be five billion NOK per year the next 10 years. In 2012 the total book value in the distribution level for the industry is 46 billion NOK.

NVE has chosen three outputs in the model; number of customers, length of high voltage network (lines and cables) and number of substations. These three variables will capture the main tasks of the DSO. The number of customers is a proxy for demand of energy, and the length of the high voltage network⁸ is a proxy for transport distance. The number of substations is a proxy for the distribution of demand. These outputs together will capture the difference between supplying 100 customers at the end of a 1000 km overhead line or to have 100 customers spread all along 1000 km of overhead lines.

14.4.1.3 Constant Returns to Scale: CRS

In the DEA-model NVE assumes constant returns to scale (CRS). This is often used in incentive regulation because even if the true technology is variable returns to scale (VRS), using CRS gives incentives to change the scale of the firm to optimal size. Size is regarded as a choice of the company.

14.4.1.4 Average Data in the Frontier

The revenue caps are calculated yearly, and the benchmarking analysis is also updated with new data every year. NVE has experienced that some of the cost elements may have considerable variations from year to year, and this may affect which companies become peers. We believe it is an advantage to apply yearly data and yearly analyses; by this changes in costs will influence the revenue relatively quickly. The industry interprets frequent variations in peer units and the frontier as the model being unstable, unreliable and unpredictable. This undermines the industry's trust in the regulatory model. NVE still prefers to analyze yearly data, but regards it is an advantage to have a more stable frontier. NVE therefore calculates the frontier as an average of data over 5 years where each company is evaluated with yearly data against this frontier. This also gives incentives for the peer companies to improve their efficiency. If they can improve their performance compared to their 5-year historical average, they can achieve a DEA score higher than 1.

Measuring companies against a frontier of constructed data, will calculate norms that may not be realistic or have an intuitive interpretation. But in practice this is a negligible problem in the Norwegian model since the norms are calibrated in stage 3.

⁸Low voltage grid is highly correlated with both customers and network stations and is not needed in the model.

14.4.2 Stage 2: Correction for Operational/Environmental Environments

It is important for a regulator that the companies find the regulation model reasonable. Most companies will claim that their company is very special and therefore cannot be compared to other companies. In a country like Norway it is reasonable to claim that companies operate under different geographical environments; it is a long country with coasts, mountains and forests. It is therefore important to consider differences in relevant operational environments in the benchmarking model. In Norway, it is also stated in the legal regulation that this issue has to be addressed.

NVE has derived geographic variables by employing GIS-analysis. The basis for this analysis is data containing the geographical coordinates of the network for each DSO. We combine the geographical network data with several thematic maps, which describe the conditions in which the network is located. By applying this technique, we have produced numerous environmental variables that we tested in the model. We have also tested numerous structural variables that describe the conditions of a company. The variables that were tested had to be based in theory or from inputs from the industry.

Many of the geographical variables are strongly correlated, which will cause problems in a linear regression model. To be able to include more aspects of a geographical condition, NVE applied factor analysis on some of the most correlated variables, creating two different composite variables. One composite variable includes wind speed, distance to coast, number of supplied islands and share of sea cables in the high voltage network. This is a composite variable that describes typical coastal environments. The other is a combination of installed capacity of distributed generation, average slope of terrain and share of overhead high voltage lines through deciduous forest. This composite variable typically describes environments in the western part of Norway. Altogether, there are five geographical variables, or Z-variables, in the model, the two composite variables described above, share of underground cables (city environments), share of overhead high voltage lines through coniferous forest (cost related to forest clearing along lines) and distance to road (network availability).

Different methods have been suggested for adjusting the DEA scores for differences in environmental variables, and Coelli et al. (2005) describe some of them. NVE has used a two-stage procedure since 2010. In the current model we regress DEA-scores from the first stage⁹ on the Z-variables. NVE improved this regression stage from 2013 so such that the independent variables are not the Z-variables themselves, but the *difference* in the Z-variable for the DSO itself and its shadow company from stage 1 (Amundsveen et al. 2014). To calculate the shadow

⁹We corrected the DEA results for bias using bootstrapping. This approach is described in Edvardsen (2004), and meets some of the criticism of serial correlation of DEA-scores by Simar and Wilson (2007).

company's Z-variables, we apply the weights from DEA.¹⁰ When we have adjusted the DEA results in stage 2, we multiply the DEA result with the company's cost base to calculate the cost norm.

14.4.3 Stage 3: Calibration of the Cost Norm

When the cost norms are calculated in the second stage, only the most efficient companies will have a cost norm that equals (or is larger than) their cost base. There are several factors that may limit companies to achieve a reasonable rate of return in this model, and the main factors are the use of book values which leads to delayed cash flows and uncertainty of the results related to measurement error and lack of comparability. NVE's response to this has been to calibrate the cost norm in stage 3 so that the sum of cost norms equals the sum of cost in the industry. This implies that the industry as a whole will receive the regulatory rate of return calculated by WACC. A company with an average DEA result will receive the regulatory rate of return (RoR), a company with higher than average DEA result can receive higher RoR, a company with lower than average DEA result will receive lower RoR.

In the calibration of the cost norms, we distribute the difference between cost base and cost norms based on each company's share of the RAB. Using the RAB as a distribution factor reduces the problem of the age effect in stage 1. In stage 1, a high RAB is unfavourable; in stage 3 it is favourable.

In this stage some of the uncertainty that follows the use of any model, is reduced. One may say that the regulator gives all the inefficiency in the industry back to the companies, but since they share a given size of the total revenue caps (decided by their result in the benchmarking model) and the analysis is repeated every year, there are strong incentives for each company to reduce costs. In order to maintain a given level of RoR a company has to keep up with the development of the "average company". The large number of the companies limits the effects of cartelisation.

14.5 Discussion and Conclusion

Weyman-Jones (2006) identified four key issues where regulatory application of efficiency and productivity analysis in practice has not kept up with the development of the theoretical and empirical methodology. These key issues are choice of methodology, sample size, specification of models (variables) and translation from results to regulatory mechanism.

¹⁰If λ_{ij} denotes the weight of DSO j on the reference set of DSO i . Then $\phi_{ij} = \lambda_{ij}x_j/\lambda_i x_i$ is DSO j 's share of the inputs for the target unit of DSO i .

We find that especially the last point is a key issue in the design of a benchmarking model; how the efficiency results will enter the regulatory model for revenues. It is important to recognize that the regulatory model is more than just the economic regulation; it also consists of direct regulations. Weyman-Jones points out that there is no settled procedure for generating price base, cost base or X-factors based on results from productivity and efficiency analysis. This may be because the legislations and indirect regulations in the different countries have developed differently during the years, and that the overall goal of the regulation may differ across countries, thus how benchmarking fits into the rest of the regulation will also be different. Both Norway and the UK have a long history of regulation of this sector, but the regulatory frameworks in the two countries are quite different. Institutional and structural factors may serve as the main explanation for this fact.

In the Norwegian regulation, the results from the benchmarking model have an important role, but more indirectly as a mean to improve efficiency. The calibration mechanism in the third stage of the benchmarking model ensures that the industry as a whole has all their costs covered in total revenue caps and a regulatory rate of return on their capital. The results from the benchmarking model are used to redistribute revenue between the companies in a zero-sum game, and in our view the application of yardstick competition over time will reveal improved efficiency gain even at the industry level. This will benefit the customers by reductions in tariffs.

Since 1998, NVE has applied different applications of DEA in the benchmarking. Differences in climate and topography have over the years been one of the main drivers in the regulatory model development. As we described in Sect. 14.4, different approaches have been applied. NVE's current methodology is a second stage regression model, which is quite popular but has also criticised, see Amundsveen et al. (2014) for references. Although NVE has started to evaluate newer models like StoNED, a change from DEA to an alternative model will not depend on theoretical arguments alone. In addition to statistical properties, the model should be intuitive, robust, and even understandable for utility managers and owners. Thus, there certainly are trade-offs in choice of model as well. As long as the results from a more advanced and theoretically correct model do not deviate too much from a simpler model, it may be better to use the simpler model.

As for sample size, the critique seems to be that the regulator does not exploit all the information available when they only use national cross sectional data. NVE thinks it is an advantage to use yearly data in order for the incentives to work rapidly; the companies can receive the gains from efficient decisions relatively quickly. On the other hand, NVE thinks we exploit some extra information when using 5 years data in the frontier of the DEA. The fact that the peers are also measured against these historical data, give the peers incentives to improve their productivity and therefore this mechanism gives incentives for the industry to become more efficient over time. As for international studies, NVE has experienced that it takes effort to make data comparable across different regulatory jurisdictions. Since Norway has relatively many companies to compare on a national level, the need for international cooperation is not urgent on DSO level. On TSO level, however, NVE has participated in international studies.

As for the choices of variables, the valuation of capital is a good illustration of the balancing of incentives that goes through the whole model. Using book values in the cost norms creates an age effect. A strong age effect gives disincentives for investments. A weaker age-effect, however, may serve as an incentive to utilize the capital for as long as possible before reinvesting. Our analyses in NVE (2011) show that the use of book values together with the calibration mechanism in stage 3 is the best solution to give the right incentives for investments. In addition, NVE allows the companies to include capital costs from investments in their tariff base without time lag, and the capital is not benchmarked during the two first years.

There is a difference between a regulatory application of benchmarking and the theoretical and empirical research. This applies to all of the four key issues that Weyman-Jones points out. In our view this is due to the fact that regulators have a wider set of goals and considerations than just applying performance measurement for improved efficiency. Based on our experience, a successful regulatory model has to find a reasonable balance between incentives for efficiency, quality of service and investments. It is also important with a reasonable distribution of efficiency gains between companies and their customers. Further, the overall total effects in the regulatory model are more important than for example the application of the ideal text-book model. It is also important to recognize that it is limited what economic regulation alone can achieve. Therefore it is crucial to apply direct regulations that define rights and obligations. Other factors like data availability and legal issues may also prevent regulators from applying text-book solutions. Due to asymmetric information, the regulatory model should reward companies that choose the optimal solutions. Last but not least, the regulator has to convince stakeholders to trust a long term sustainable regulatory framework that gives the possibility to earn a reasonable rate of return and that will be adapted to future changes in constraints and environments.

References

- Amundsveen R, Kvile HM, Kordahl OP, Langset T (2014) Second stage adjustment for firm heterogeneity in DEA: a novel approach used in regulation of Norwegian electricity DSOs. In: Emrouznejad A, Bunker R, Doraisamy SM, Arabi B (eds) Recent developments in data envelopment analysis and its applications, Proceedings of the 12th International Conference of DEA, Kuala Lumpur, April 2014
- Coelli TJ, Rao DSP, O'Donnell CJ, Battese GE (2005) An introduction to efficiency and productivity analysis, 2nd edn. Springer, New York
- Edvardsen DF (2004) Efficiency of Norwegian construction firms, four essays on the measurement of productive efficiency. Doctoral thesis, University of Gothenburg
- Haney AB, Pollitt MG (2012) International benchmarking of electricity transmission by regulators: theory and practice. EPRG working paper 1226
- Kittelsen SAC (1994) Effektivitet og regulerings i norsk elektrisitetsdistribusjon. SNF-report 3/1994
- Kuosmanen T (2012) Stochastic semi-nonparametric frontier estimation of electricity distribution networks: application of the StoNED method in the Finnish regulatory model. Energy Econ 34:2189–2199. doi:[10.1016/j.eneco.2012.03.005](https://doi.org/10.1016/j.eneco.2012.03.005)

- Langset T, Syvertsen SC (2013) A new WACC model in the regulation of the Norwegian electricity network operators. Paper presented at the 1st Conference of the RCEM & ICSTF, ESCP Europe Business School, London
- Ministry of Petroleum and Energy (2014) Et bedre organisert strømnett. Report from an expert commission, Oslo
- NVE (2009) Endringer i normkostnadsmodellen for distribusjonsnettet. Memorandum from 22 October 2009
- NVE (2011) Alderseffekter i NVEs kostnadsnormer—evaluering og analyser. NVE-report 21/2011
- Shleifer A (1985) A theory of yardstick competition. *Rand J Econ* 16(3, autumn):319–332
- Simar L, Wilson PW (2007) Estimation and inference in two-stage, semi-parametric models of production processes. *J Econ* 136(1):31–64. doi:[10.1016/j.jeconom.2005.07.009](https://doi.org/10.1016/j.jeconom.2005.07.009)
- Waddams C (2000) Efficiency and productivity studies in incentive regulation of UK utilities. *Revista de Economia del Rosario* 3(2)
- Weyman-Jones T (2006) Efficiency and productivity analysis in regulation and governance. 4th NAPW Plenary Paper, New York

Chapter 15

Limitations of the Approximation Capabilities of the Translog Model: Implications for Energy Demand and Technical Change Analysis

Sourour Baccar

Abstract In this paper, we evaluate the capacity of the Translog cost share model to approximate the producer's true demand system and introduces two non-linear functional forms, which have been achieved by altering and extending the standard quadratic logarithmic Translog model. The extensions have additional desirable approximation properties with respect to output and time variables, and thus allow more flexible treatments of non-homothetic technologies and non-neutral technical change than those provided by the standard Translog. The performances of the three models are assessed (1) on theoretical ground, by the size of the domain of regularity, (2) on their ability to provide plausible estimates of the economic and technological indicators being measured and finally (3) on their reliability in fitting input shares, input-output ratios and unit cost. The most important finding is that the standard model exhibits some weakness in fitting. We show via a series of experiments that those shortcomings are due to a lack of flexibility of the logarithmic model. The estimation results obtained with the new extended model are more satisfactory and promising.

Keywords Translog • Cost function • Flexibility • Technical change modeling • Energy demand • Reliability of fit

JEL code: C51, C52, L60, O30

S. Baccar (✉)

Department of Mathematics and Statistics, University of Sfax, Sfax, Tunisia

e-mail: ss.baccar@gmail.com

15.1 Introduction

The developments in the applications of duality theory to economic analysis has stimulated extensive research on the econometric specification of factor and consumer-demand systems.¹ Much of this work, pioneered by Diewert (1971), has focused on an attempt to search for general functional forms for cost (or indirect utility) functions which can allow derivation of factor (or consumer) demand systems that will reliably approximate systems generated by a broad range of cost (or indirect utility) functions, and can provide the capability to attain an arbitrary set of elasticities. As a result, several families of parametric models called “flexible functional forms” have been proposed, which supplied attractive alternatives to previous inflexible models such as the popular Cobb-Douglas and CES functional forms.² The most likely contribution of these models lies in the fact that they enable us to infer the structure of production (or preferences) without many prior restrictions about the economic indicators being measured (such as values and signs of elasticities of substitution) while retaining direct contact with underlying economic theory, and to test many important hypotheses in production and consumption theory. Thus, the introduction of flexible functional forms has made it possible to overcome the limitations of parametric forms based on constant elasticities of substitution criticized by Uzawa (1962).³

The specification of a flexible functional form is based on approximation theory. Generally, this involves the consideration of a class of series expansion which can be used to generate accurate approximations to some unknown function (satisfying the appropriate regularity conditions). For instances, one has functional forms based on Taylor series expansion (translog [Christensen et al. 1971, 1973], generalized Leontief [Diewert 1971] and symmetric generalized McFadden [McFadden 1978 and Diewert and Wales 1987]); on Box-Cox transformation (generalized Box-Cox [Berndt and Khaled 1979]); on Fourier series expansion (Fourier flexible form [Gallant 1981]); on Laurent series expansion (Minflex Laurent [Barnett 1983a,b]

¹The duality theory provides alternative equivalent representations of a producer's technology (production, cost, profit, revenue, or distance function) or of a consumer's preferences (direct utility, indirect utility, inverse indirect utility, or expenditure function). Under certain regularity conditions, given one of these functions the other can be determined and completely characterized. See Diewert (1974, 1982) for excellent surveys of the applications of duality theory in different areas of economics. See also Fuss et al. (1978) for a comprehensive treatment of the application of duality to production theory.

²The class of flexible functional forms was originally defined by Diewert (1971, 1974) to be the class of forms capable of attaining the level of the “true” function and of all its first and second order partial derivatives at some point. Barnett (1983a,b) later calls this local flexibility. Gallent (1981) distinguish between two concepts of flexibility: Diewert or Sobolev.

³Uzawa (1962) proved that it is impossible for any functional form that exhibits constant elasticities of substitution to provide simultaneously the capability to attain an arbitrary set of elasticities. See Jorgenson (1986) and Lau (1986) for excellent surveys of the literature on specification of functional form and econometric modeling of producer behavior.

and symmetric generalized Barnett [Diewert and Wales 1987]); and on Muntz-Szasz series expansion (Asymptotically Ideal Model [Barnett and Jonas 1983]). The list of models available is still growing. However, although efforts to develop new flexible forms enrich the family of specification alternatives, most applications with these functions have favored Diewert's class of locally flexible forms, or equivalently functional forms achieving second order approximations to any arbitrary twice differentiable function $F^*(x)$ at a given point x^* .⁴ Most of the currently available elements of that class were generated by truncating the series expansion at second order terms. Unfortunately, these functional forms rely on an intrinsically local notion of flexibility and may behave poorly over a finite region as has been shown in a number of empirical applications and Monte Carlo experiments.⁵ Another body of literature has dealt with globally flexible functional forms, representations that are typically infinite order series expansion that must be truncated at some point. Such functional forms include the Fourier flexible form of Gallant (1981) and the Asymptotically Ideal Model of Barnett and Jonas (1983), which approximate the true function in the so called Sobolev norm, have the ability of achieving global approximations to any arbitrary continuous function and its partial derivatives.⁶ Unfortunately, while very general, these functional forms are not parsimonious (in terms of number of parameters) and more cumbersome to implement empirically.⁷

At this end, still the functional form that has proven the most popular and widely used in empirical economic research is the translog model of Christensen et al. (1971, 1973), which can be interpreted as a logarithmic second-order Taylor series expansion of an arbitrary twice differentiable function $F^*(x)$ in the neighborhood of $x^* = (1, \dots, 1)^T$.

⁴According to Diewert's (1971, 1974) definition, a function $G(x)$ is a second order approximation to an arbitrary function $F^*(x)$ at a given point x^* if the level, the first and second order partial derivatives of these two functions are equal at these point. Lau (1974) distinguished between two definitions to the concept of 'second order approximation' that are usually used in economic literature, by referring to Diewert's definition as 'second order differential' approximation and to Christensen, Jorgenson and Lau's (1973) definition as 'second order numerical' approximation. Barnett (1983a,b) later proved that flexibility in the sens defined by Diewert is necessary and sufficient for a function to satisfy the mathematical definition of a local second order approximation. Furthermore, a second order Taylor series approximation is flexible in that sense.

⁵See, e.g., Caves and Christensen (1980), Guilkey and Lovell (1980), Guilkey et al. (1983), Gallant (1981), Barnett (1983a,b), Barnett and Lee (1985), Barnett et al. (1985), and Diewert and Wales (1987).

⁶As has been noted by Gallant and Golub (1984), a Sobolev-flexible form can endow a parametric methodology with a semi-nonparametric property.

⁷A functional form is parsimonious if it provides a second order approximation using a minimal number of parameters. Barnett (1985) refer to that property, first defined by Diewert (1971), as the *minimality property*. The term "parsimonious" is due to Fuss et al. (1978). As these authors noted, an excessive number of parameters exacerbate problems of multicollinearity. Furthermore, when the sample is small, excess parameters mean a loss of freedom and hence a loss in the precision of estimation.

This preference towards the translog may be justified by a variety of reasons. The use of the logarithmic metric to all right side variables and to the dependent variable of the approximating function allows for global homogeneity via simple parametric restrictions, which can be readily imposed prior to estimation. In addition, its log-quadratic structure with its linearity in the parameters permits the effortless derivation of elasticities (at least for long-run equilibrium models) while regression coefficients have an intrinsic and intuitive economic interpretation.⁸ Finally, this specification is used as a benchmark since it has been extensively used over the years. This phenomenon is self-reinforcing since the researcher is strongly induced to compare directly the empirical results obtained to some reference group according to the nature of the data, the approach pursued, modeling characteristics, etc.

The main subject of this paper is to discuss the ability of the translog model to approximate the producer's true demand system. Then, instead of choosing a single functional form as an approximation to the true cost function, along with the translog (henceforth *TL*) two other commonly used flexible forms have been confronted: the generalized Leontief (henceforth *GL*) due to Diewert (1971) and the symmetric generalized McFadden (henceforth *MF*) due to McFadden (1978), and Diewert and Wales (1987).⁹ While maintaining an identical data set and retaining the same assumptions about the technology, an empirical comparison of alternative flexible functional forms is attempted. This comparative study casts doubt on the reliability of the translog model. That is, the precision of fit of translog may seem satisfactory if one limits oneself on scrutinizing the predicted cost share of each input, that are the dependent variables of the standard regression system of translog, but it is no longer so if the predicted values for unit cost and input-output ratios are also scrutinized. Meanwhile, the results stemming from the two other forms *GL* and *MF* are satisfactory in terms of the goodness of fit, when the three preceding criteria are gathered. These shortcomings of *TL* in fitting factor's demand and unit cost, in comparison to *GL* and *MF*, seem independent to the hypothesis underlying

⁸The use of the translog for short-run studies in production and demand analysis entails a significant shortcoming in that the full equilibrium level(s) of fixed input(s) cannot be derived in analytical closed form(s), but instead must be calculated using iterative numerical techniques. However, some researchers have reported difficulties in obtaining numerical convergence with the translog variable cost function (assuming temporary equilibrium) and thus with computing estimates for long-run elasticities and capacity utilization (see, e.g., Brown and Christensen 1981; Berndt and Hesse 1986; Baccar 2006). Moreover, problems may rise with frameworks based on cost of adjustment or time to build dynamic models.

⁹The three functional forms considered belong to the generalized quadratic family of locally flexible functional forms and can be interpreted as second order Taylor series expansion about a point in powers of $\ln x$, $x^{1/2}$ and x respectively. They share the common characteristics of linearity in parameters and the “ability” of providing second-order approximations to an arbitrary twice continuously differentiable function at a point. All of these forms can be viewed as limiting or special cases of a more general quadratic form: the generalized Box-Cox due to Berndt and Khaled (1979).

the adjustments of inputs and to the structure of the technology. It reveals some weakness of the *TL* specification itself (see Baccar 1995a).

Surprisingly, this problem has never been treated, not even signalled, to my knowledge, in the vast empirical literature based on translog specifications. In fact, most if not all previous studies concerning empirical comparisons of flexible functional forms employed two principal criterions as the basis for evaluating the performances of such forms: (1) the theoretical performances by the size of their regular region; and (2) the economic performances by the ability of each form to provide plausible estimates of the economic indicators being measured. Unfortunately, when flexible forms parameters are estimated from time series data, the curvature conditions implied by economic theory are often violated, while imposing these conditions globally prior to estimation could significantly restrict the flexibility of most flexible forms. While the second criterion relies too much on intuition since the true values of economic indicators are unobservable a priori. Additional criterion which have been neglected in the published literature comparing the performances of flexible forms is proposed: the ability of the approximating form to “fit” the endogenous variables of the model (i.e. input shares, input-output ratios and unit cost in the case of a cost function) which are observable and directly available.

The contribution of my work is twofold. First, to reveal the weakness of the translog model. Second, and even more important, to locate the origin of those shortcomings following two lines of research. The first line of research consists in exploring some aspects of the modeling essentially linked to the use of the logarithmic metric which associates a *multiplicative* stochastic error with the estimated cost function and imposes an *exponential* representation of technical change parameters. Two modeling characteristics are investigated: first, alternative stochastic specifications (additive or multiplicative), and second, alternative specifications of the disembodied technical change (exponential or linear). For this purpose, I develop two nonlinear versions of the translog cost function with exponential and linear technical progress, respectively. The pure effects of alternative stochastic specifications and technical change modeling are analyzed. The second line of research comes back to the parametrization of the estimated model. It consists in exploring different degrees of flexibility of the *log-quadratic* version, by varying the restrictions imposed on the disembodied technical change and scale economies parameters.

This framework is based on a long-run equilibrium approach with variability of all factor inputs including capital stock. This choice is justified by two principal reasons: (1) it is the simplest and the more often used in the comparison between the performances of the functional forms; (2) to facilitate understanding of the dynamic translog specifications, we will first review the static model.

15.2 Theoretical Background

Suppose that the technology of a firm (or an industry) can be represented by a well-behaved production function $f : y = f(x, t)$; where y is the maximum output level obtainable in period t from the N -dimensional vector of variable inputs $x \equiv (x_1, \dots, x_N)^T \geq 0_N$, and t is a time variable incorporated in f as a proxy for the level of technological change.¹⁰ f satisfies certain regularity conditions, namely, $f(0_N, t) = 0$, f is nonnegative real valued, non-decreasing, strictly quasi-concave and twice continuously differentiable in x . Letting $p \equiv (p_1, \dots, p_N)^T \gg 0_N$ be the N -dimensional vector of input prices. Assume that firm (or industry) is price taker and attempts to minimize the cost of producing output level y . Duality theory states that technology can be equivalently represented by a total cost function C dual to the production function f , defined as the solution to the following constrained minimization problem:

$$C \equiv C(p, y, t) \equiv \min_x \{p^T x : f(x, t) \geq y\}, \quad (15.1)$$

where $p^T x$ is the inner product $\sum_{i=1}^N p_i x_i$.

The optimal value function $C(p, y, t)$, defined by the cost-minimizing behavior problem (15.1), gives the minimum cost of achieving output level y during a period of time as a function of input prices p , output y , and time t . The time variable t is added to the exogenous variables included in C to enable the measurement of technical change.¹¹

Given the assumptions made on f , duality theorems establish that C satisfies certain properties, often referred to as regularity conditions: the cost function must be nonnegative real valued for all $p \gg 0$ and $y \geq 0$, nondecreasing in p and y , linearly homogeneous and concave in p .¹² For mathematical convenience, we shall assume throughout that C is at least twice continuously differentiable in its $N + 2$ arguments and involves $N + 2$ first and $(N + 2)(N + 3)/2$ second partial derivatives. The above consistency properties of the cost function may be mathematically expressed in terms of C and its first and second derivatives as summarized in Table 15.1.

¹⁰Notations: $x \geq 0_N$ means each element of the N -dimensional vector x is nonnegative, and $x \gg 0_N$ means that each element of x is positive.

¹¹This cost function is derived under the assumption that the observed production technology is instantaneously in full static equilibrium. Thus, all inputs are assumed to adjust fully to their long-run equilibrium levels within one sample period. A variable (or restricted) cost function can also be specified to allow for short-run fixity of some inputs in studies based either on partial-static or dynamic equilibrium models. See Baccar (2006) for further discussions in this topic.

¹²For extensive discussions on the various duality relationships between production and cost functions, the reader is referred to, for example, Shephard (1953, 1970) or Diewert (1971, 1974, 1982).

Table 15.1 Theoretical consistency properties of the cost function

1. Domain	$C(p, y, t) \geq 0$, for all $p \gg 0_N$, $y > 0$ (nonnegativity); $C(p, 0, t) = 0$ (no fixed costs)
2. Monotonicity	$C(p, y, t)$ is nondecreasing in p and y . Then $\frac{\partial C}{\partial p_i} \geq 0$, for all i and $\frac{\partial C}{\partial y} \geq 0$
3. Homogeneity	$C(p, y, t)$ is linear homogeneous in p . Hence Euler Theorem yields ^a $\sum_{i=1}^n p_i \frac{\partial C}{\partial p_i} = C$ (2.a), $\sum_{i=1}^n p_i \frac{\partial^2 C}{\partial p_i \partial p_j} = 0$, for all j (2.b) $\sum_{i=1}^n p_i \frac{\partial^2 C}{\partial p_i \partial y} = \frac{\partial C}{\partial y}$ (2.c), $\sum_{i=1}^n p_i \frac{\partial^2 C}{\partial p_i \partial t} = \frac{\partial C}{\partial t}$ (2.d)
4. Symmetry	$C(p, y, t)$ is twice continuous differentiable, young theorem implies $\frac{\partial^2 C}{\partial p_i \partial p_j} = \frac{\partial^2 C}{\partial p_j \partial p_i}$, for all i, j i.e. that the Hessian matrix must be symmetric
5. Concavity	$C(p, y, t)$ is concave in p if $\left[\frac{\partial^2 C}{\partial p_i \partial p_j} \right]$ is a negative semidefinite matrix ^b , i.e. that the Hessian matrix must be negative semidefinite

^aThe homogeneity restrictions (2.a), (2.b), (2.c) are known us the adding-up condition, the Cournot aggregation conditions, and the Engel aggregation condition, respectively

^bConcavity and twice continuous-differentiable of $C(p, y, t)$ imply that $\left[\frac{\partial^2 C}{\partial p_i \partial p_j} \right]$ is negative semidefinite matrix (see Rockafellar 1970; Diewert et al. 1981)

However, under differentiability the cost function possesses a very useful property for theoretical and empirical applications of duality theory:

if C satisfies regularity conditions above and, in addition, is differentiable in p , then by Shephard's Lemma,

$$x_i = x_i(p, y, t) = \frac{\partial C(p, y, t)}{\partial p_i}, \quad \forall i \quad (\text{Shephard Lemma}) \quad (15.2)$$

where $x_i(p, y, t)$ is the cost-minimizing demand for the i th input needed to produce y units of output given positive input prices $p \gg 0_N$.¹³ Then, $x_i(p, y, t)$ must be continuous and zero homogeneous in p and nonincreasing in p_i , for all i .

Thus, differentiation of the cost function with respect to input prices yields the system of cost-minimizing input demand functions, $x(p, y, t) = \nabla_p C(p, y, t)$, where $\nabla_p C(\cdot) \equiv [\partial C(\cdot)/\partial p_1, \dots, \partial C(\cdot)/\partial p_N]^T$ is the gradient of $C(\cdot)$ in p .

¹³Although Shephard's Lemma has been stated and proven by many authors, the first complete proof was provided by Shephard (1953). See Diewert's surveys (1974, 1982) for further discussions on this topic. For a detailed proof of the theorem, the reader is referred to Diewert (1971).

Usually, at a point of cost minimization, from Shephard's Lemma, we get the cost-minimizing input demand system by partially differentiating the cost function with respect to input prices, and after dividing through by output the system of input-output ratios $(a_1, \dots, a_N)^T$; or the cost-minimizing input cost share system $(S_1, \dots, S_N)^T$ by partially differentiating the logarithm of the cost function with respect to the logarithm of input prices.¹⁴ More specifically,

1. *Direct specification:* $C \equiv C(p_1, \dots, p_N, y, t)$

$$a_i = a_i(p, y, t) = \frac{x_i}{y} = \frac{1}{y} \frac{\partial C}{\partial p_i} \quad \forall i, \quad (15.3)$$

2. *Logarithmic specification:* $\ln C \equiv \ln C(\ln p_1, \dots, \ln p_N, \ln y, t)$

$$S_i = S_i(p, y, t) = \frac{p_i x_i}{C} = \frac{p_i}{C} \frac{\partial C}{\partial p_i} = \frac{\partial \ln C}{\partial \ln p_i} \quad \forall i, \quad (15.4)$$

where

$$\sum_{i=1}^N S_i = 1 \quad (\text{1 adding-up restriction}). \quad (15.5)$$

Upon specification of a functional form for the cost function, the estimation of the cost-minimizing input demand system, or of the log cost function in conjunction with $N - 1$ corresponding input cost share equations allows for approximations of various economic indicators of interest such as elasticities of substitution, returns to scale, and bias of technical change.

The evaluation of these indicators yields many useful insights, especially in terms of the patterns of input substitutability, the implications of scale economies, and the effect of technological change in a given production system.

The most widely used measure of input substitutability is the Allen partial elasticity of substitution σ_{ij} between inputs i and j . Uzawa (1962), showed that σ_{ij} can be defined in terms of the cost function as

$$\sigma_{ij} \equiv \frac{CC_{ij}}{C_i C_j} \quad (15.6)$$

where $C_i = \partial C / \partial p_i$ and $C_{ij} = \partial^2 C / \partial p_i \partial p_j$. From Shephard's Lemma it follows that $C_i = x_i = S_i \cdot C / p_i$, then $C_{ij} = \partial x_i / \partial p_j = C(\partial^2 \log C / \partial \log p_i \partial \log p_j - \delta_{ij}S_i + S_i S_j) / p_i p_j = \partial x_j / \partial p_i = C_{ji}$. Hence, σ_{ij} may be written as

¹⁴From Shephard's lemma, $\partial C / \partial p_i = x_i$. Consider a logarithmic transformation of the cost function, we find that the logarithmic derivatives of the cost function are input cost shares.

$$\sigma_{ij} = \frac{C}{x_i x_j} \frac{\partial x_i}{\partial p_j} = \frac{C/y}{a_i a_j} \frac{\partial a_i}{\partial p_j} = \frac{1}{S_i S_j} \left[\frac{\partial^2 \log C}{\partial \log p_i \partial \log p_j} - \delta_{ij} S_i + S_i S_j \right] \quad (15.7)$$

where δ_{ij} is the Kronecker delta. Thus σ_{ij} can be regarded as the normalized response of the i th input cost-minimizing demand to a change in the price of the j th input, where the normalization is chosen so that $\sigma_{ij} = \sigma_{ji}$ and so that σ_{ij} is invariant to changes in the scale measurement of inputs.

The own and cross price elasticities of demand, ε_{ij} , are conventionally defined as

$$\varepsilon_{ij} = \frac{\partial \ln x_i}{\partial \ln p_j}. \quad (15.8)$$

These elasticities are analytically related to the Allen elasticities of substitution by

$$\varepsilon_{ij} = S_j \cdot \sigma_{ij} \quad (15.9)$$

where inputs i and j are Allen substitutes if $\varepsilon_{ij} > 0$, ‘independents’ if $\varepsilon_{ij} = 0$, and Allen complements if $\varepsilon_{ij} < 0$.

Notice that the cross-price elasticities of demand are not symmetric. Thus, even though $\text{sing}(\varepsilon_{ij}) = \text{sing}(\varepsilon_{ji}) = \text{sing}(\sigma_{ij})$, in general $\varepsilon_{ij} \neq \varepsilon_{ji}$ for $i \neq j$.¹⁵ However, it is easy to show that $\varepsilon_{ij} = (S_j/S_i) \cdot \sigma_{ji}$. The own price elasticities of demand ε_{ii} measures the change in i th input use in response to its own price variations, when output and the prices of all other input remain constant. The concavity property of the cost function requires that $\varepsilon_{ii} \leq 0$ for all i . Finally, the Cournot aggregation conditions ($\sum_{j=1}^N p_j C_{ij} = 0$, $\forall i$) implied by linear homogeneity of C in p yield the following restrictions on the price elasticities

$$\sum_{j=1}^N \varepsilon_{ij} = 0 \Leftrightarrow \varepsilon_{ii} = - \sum_{j \neq i} \varepsilon_{ij}, \quad (\text{N restrictions}) \quad (15.10)$$

or equivalently,

$$\sum_{j=1}^N S_j \sigma_{ij} = 0 \Leftrightarrow \sigma_{ii} = - \sum_{j \neq i} \frac{S_j}{S_i} \sigma_{ij}, \quad (\forall i) \quad (15.11)$$

The degree of returns to scale, say η , is defined as the proportional change in output attributable to a proportional change in all inputs with time fixed. At cost minimizing points, η is the reciprocal of the elasticity of cost with respect to output. More specifically,

¹⁵In both measures, σ_{ij} and ε_{ij} the quantities of all inputs are allowed to adjust in response to a change in p_j , when output is held constant.

$$\eta = \varepsilon_{Cy}^{-1} \quad \text{where} \quad \varepsilon_{Cy} = \frac{\partial \ln C}{\partial \ln y} \quad (15.12)$$

If $\varepsilon_{Cy} <, =, > 1$, then the dual production function exhibits locally increasing, constant, or decreasing returns to scale, respectively.

The proportional change in the i th cost-minimizing demand in response to a proportional change in output is given by

$$\varepsilon_{iy} = \frac{\partial \ln x_i}{\partial \ln y} \quad (15.13)$$

The Engel aggregation condition implies

$$\varepsilon_{Cy} = \sum_{i=1}^N S_i \cdot \varepsilon_{iy} \quad (\text{1 restriction}) \quad (15.14)$$

The rate of technical change, say T_C , is typically defined as the rate of change of output over time, holding all inputs constant: $T_C = \partial \ln y / \partial t$. As has been shown by Ohta (1974), this measure can be obtained from the cost function as

$$T_C = -\frac{\varepsilon_{Ct}}{\varepsilon_{Cy}} \quad (15.15)$$

where $\varepsilon_{Ct} = \partial \ln C / \partial t$ is the rate of change in total cost over time that is not attributable to changes in p or y . Often, ε_{Ct} is referred to as the *rate of cost diminution* due to disembodied technical change. When the technology exhibits constant returns to scale, $\varepsilon_{Cy} = 1$, hence from (15.15) $T_C = -\varepsilon_{Ct}$. Thus with constant returns to scale the primal rate of technical change is the negative of the dual rate of cost diminution.

The bias of technical change can also be categorized. The measure of the direction of biased technical change with respect to the i th input can be represented by (1) the rate of change in the demand of the i th input over time, say ε_{it} , or by (2) the rate of change in the i th input cost share over time, say I_{bi} . The corresponding expressions of these two measures are respectively:

$$\varepsilon_{it} = \frac{\partial \ln x_{it}}{\partial t} \quad \forall i, \quad (15.16)$$

$$I_{bi} = \frac{\partial \ln S_{it}}{\partial t} = \varepsilon_{it} - \varepsilon_{Ct} \quad \forall i \quad (15.17)$$

Technical change is often said to be input- i (share- i) saving if $\varepsilon_{it} < 0$ ($I_{bi} < 0$), neutral if $I_{bi} = 0$, and input- i (share- i) using if $\varepsilon_{it} > 0$ ($I_{bi} > 0$).¹⁶

Since the cost function is to be homogeneous of degree one in p , Euler's theorem on homogeneous functions implies

$$\varepsilon_{Ct} = \sum_{i=1}^N S_i \cdot \varepsilon_{it} \quad (15.18)$$

For empirical investigations, a functional form is needed to specify the dual cost function. In this paper, three alternative versions of the translog cost function are proposed. First, the quadratic logarithmic form (that is the *standard* version of translog that appears in the literature) which approximates the natural logarithm of total cost by a quadratic function of the logarithm of input prices, the logarithm of output, and time. Second, the exponential of the quadratic logarithmic form, which is *nonlinear* in the parameters and approximates total cost by a transcendental or, more specifically, an exponential function of the logarithm of input prices, the logarithm of output, and time. Finally, a modified form of the nonlinear version of translog, that is the second *nonlinear* version, introduced here for comparative purpose. However, the modifications have additional desirable approximation properties with respect to time and output variables, and thus permit a more flexible treatment of nonhomothetic technologies and nonneutral technical change than is provided by the popular quadratic logarithmic form. These two nonlinear versions of translog enable us to analyze the pure effects of alternative technical change modeling on the behavior of the estimated elasticities. Along with the translog, two other specifications are invoked: the generalized Leontief (*GL*) and the symmetric generalized McFadden (*MF*).

15.3 The Models

15.3.1 The Standard Version of Translog (TL)

The well-known translog approximation to an arbitrary cost function that allows for nonhomotheticity and nonneutral technical change can be written as

$$\ln C(p, y, t)_{TL}$$

¹⁶Hicks defined technical change as being neutral if the marginal rate of technological substitution between each pair of inputs is independent of technical change. Thus, when technical change is Hicks-neutral all factor demands are affected equi-proportionally. Hicksian bias is defined as the change in factor-ratio (x_i/x_j) or in factor share-ratio (S_i/S_j) that is not attributable to price changes.

$$\begin{aligned}
&\equiv \alpha_0 + \sum_{i=1}^N \alpha_i \ln p_i + \alpha_y \ln y + \alpha_t t \\
&+ 0.5 \sum_{i=1}^N \sum_{j=1}^N \alpha_{ij} \ln p_i \ln p_j + \sum_{i=1}^N \alpha_{iy} \ln y \ln p_i + \sum_{i=1}^N \alpha_{it} t \ln p_i \\
&+ 0.5 \alpha_{yy} \ln y^2 + \alpha_{yt} t \ln y + 0.5 \alpha_{tt} t^2, \quad \alpha_{ij} = \alpha_{ji}, \quad \forall i, j
\end{aligned} \tag{15.19}$$

where α_0 , α_i , α_y , α_t , α_{ij} , α_{iy} , α_{it} , α_{yy} , α_{yt} , and α_{tt} are parameters to be estimated, while the symmetry of the second order parameters is a necessary and sufficient condition for the symmetry of the Hessian matrix. The cost function defined by (15.19) contains $N(N + 1)/2 + 3N + 6$ free parameters.

Applying Shephard's Lemma in its logarithmic version, the translog approximation to the system of input cost share equations derived from the log cost function (15.19) can be written as

$$S_i(p, y, t)_{TL} \equiv \alpha_i + \sum_{j=1}^N \alpha_{ij} \ln p_j + \alpha_{iy} \ln y + \alpha_{it} t, \quad \forall i \tag{15.20}$$

which are zero homogeneous in input prices. Note that the translog cost function and its associated input cost share equations are linear in the parameters.

Since a well behaved cost function must be consistent with linear homogeneity in p , the parameterized cost shares (15.20), which are the first order derivatives of (15.19) with respect to the logarithm of input prices, must sum to unity: $\sum_{i=1}^N S_i = 1$. Hence,

$$\sum_{i=1}^N \alpha_i + \sum_{i=1}^N \sum_{j=1}^N \alpha_{ij} \ln p_j + \sum_{i=1}^N \alpha_{iy} \ln y + \sum_{i=1}^N \alpha_{it} t = 1 \tag{15.21}$$

So, linear homogeneity in the vector of input prices can be imposed globally on (15.19) via the following $N + 3$ adding-up restrictions relating the parameters of (15.20)

$$\begin{aligned}
\sum_{i=1}^N \alpha_i &= 1, & \sum_{j=1}^N \alpha_{ij} &= 0, \quad \forall i \\
\sum_{i=1}^N \alpha_{iy} &= 0, & \sum_{i=1}^N \alpha_{it} &= 0
\end{aligned} \tag{15.22}$$

The reader can easily verify that incorporating the above restrictions into (15.19) reduces the number of free parameters to $N(N + 1)/2 + 2N + 3$.

The output elasticity of the cost function associated with the translog approximation defined by (15.19) and (15.22) is given by

$$\varepsilon_{Cy} = \alpha_y + \sum_{i=1}^N \alpha_{iy} \ln p_i + \alpha_{yy} \ln y + \alpha_{yt} t \quad (15.23)$$

The cost function would be homothetic if it could be written as a separable function of output and input prices: $C(p, y) = h(y) \cdot C(p)$. If the translog approximation defined by Eq. (15.19) is to be consistent globally with homotheticity, it must be true that $\alpha_{iy} = 0$ for all i . By this result, the dual production function is homogeneous only if $\alpha_{yy} = 0$.

Constant returns to scale (CRTS) occurs (so the cost function is linear homogeneous in output y) when the following $N + 2$ restrictions are satisfied by the parameters of (15.19) and (15.22)

$$1 - \alpha_y = \alpha_{yt} = \alpha_{yy} = \alpha_{iy} = 0, \forall i \quad (15.24)$$

Thus, with CRTS parameter restrictions imposed, the translog approximation to the unit cost function can be written in the simplified form:

$$\begin{aligned} \ln c(p, t)_{TL} = & \alpha_0 + \sum_{i=1}^N \alpha_i \ln p_i + \alpha_t t + 0.5 \sum_{i=1}^N \sum_{j=1}^N \alpha_{ij} \ln p_i \ln p_j \\ & + \sum_{i=1}^N \alpha_{it} t \ln p_i + 0.5 \alpha_{tt} t^2, \quad \alpha_{ij} = \alpha_{ji}, \quad \forall i, j \end{aligned} \quad (15.25)$$

where $c = C/y$ is the unit cost of production.¹⁷ Then, imposing the CRTS parameter restrictions reduce the number of free parameters to $(N + 2)(N + 1)/2$.

The overall rate of technical change derived from Eq. (15.25) is approximated by

$$T_C = -\frac{\partial \ln c(p, t)}{\partial t} = -(\alpha_t + \sum_{i=1}^N \alpha_{it} \ln p_i + 0.5 \alpha_{tt} t^2) \quad (15.26)$$

The technical change is said to be unbiased (or share neutral) if it leaves relative cost shares undistributed. So under linear homogeneity in p and y , we need only to impose the $N - 1$ additional restrictions

$$\alpha_{it} = 0, \forall i \quad (15.27)$$

If in addition $\alpha_{tt} = 0$, then technical change is assumed to occur at a constant proportional rate while the first order neutral component α_t shows the autonomous reduction in unit cost in response to improved technology.

¹⁷Under constant returns to scale, the unit cost function is independent of y .

The Allen-Uzawa elasticities of substitution derived from the translog approximation are given by

$$\sigma_{ij} = \frac{\alpha_{ij} + S_i S_j}{S_i S_j}, \quad \forall i \neq j \quad (15.28)$$

The own and cross price elasticities of demand can be expressed as

$$\varepsilon_{ij} = \frac{\alpha_{ij} - \delta_{ij} S_i + S_i S_j}{S_i}, \quad \forall i \quad (15.29)$$

where δ_{ij} is the Kronecker delta.

Before presenting the two nonlinear versions of translog, we give some remarks about the analytical structure of the standard version *TL*.

1. It is not possible to recover all the empirical implications of the cost-minimizing model of the firm (or industry) from the share system: the efficiency parameter α_0 , and furthermore the first and the second order autonomous parameters determining the rate of returns to scale (α_y, α_{yy}) and those determining the overall rate of technical change (α_t, α_n) only appear in the cost function.
2. The parameter α_{it} , which captures the effects of biased technical change on the i th input demand, can be interpreted as constant share elasticity of input i with respect to time. However, since the cost function is to be linear homogeneous in input prices, these parameters must sum to zero ($\sum_{i=1}^N \alpha_{it} = 0$). This means that the bias of technical change with respect to the price of one input is restricted to be the negative of the sum of biases technical change with respect to the prices of the other inputs ($\alpha_{it} = -\sum_{j \neq i} \alpha_{jt}$). So that, the translog approximation rules out the possibility of either positivity or negativity for all the α_{it} . This structure casts some doubt in the reliability of those parameters to obtain an accurate approximation to the direction of technical change bias with respect to each input, and this difficulty increases with the number of inputs considered.
3. The parameters α_{iy} allowing for non homotheticity can be interpreted as constant share elasticities with respect to output and must sum to zero: $\sum_{i=1}^N \alpha_{iy} = 0$.
4. With the translog cost function it is not possible to restrict attention to some parameters, e.g. the energy-related ones for studying structural changes induced by shocks on energy prices. Because of the adding-up restrictions assuring linear homogeneity in input prices, it is not possible to concentrate on the energy-related parameters alone, but there should be corresponding parameter changes also in all the other parameters of the translog cost function.

15.3.2 The First Nonlinear Version of Translog (TLE)

This *nonlinear* version of translog is from an analytical point of view strictly equivalent to the *standard* quadratic logarithmic version, the difference is laying in the stochastic specification. Its general formula allowing for nonhomotheticity and for nonneutral technical change is obtained by taking antilog of (15.19). A little manipulation yields

$$C(p, y, t)_{TLE} \equiv C(p) \cdot \exp(\alpha_t t + \alpha_{tt} t^2) \cdot y^{(\alpha_y + 0.5\alpha_{yy} \ln y + \alpha_{yt} t)} \quad (15.30)$$

$$\cdot \prod_{i=1}^N p_i^{(\alpha_{iy} \ln y)} \cdot \exp(t \sum_{i=1}^N \alpha_{it} \ln p_i),$$

where the function $C(p)$ is defined by:

$$C(p) \equiv \alpha_0 \cdot \prod_{i=1}^N p_i^{(\alpha_i + 0.5 \sum_{j=1}^N \alpha_{ij} \ln p_j)}, \quad \alpha_{ij} = \alpha_{ji} \quad \forall i, j, \quad (15.31)$$

and $\alpha_0, \alpha_i, \alpha_y, \alpha_t, \alpha_{ij}, \alpha_{iy}, \alpha_{it}, \alpha_{yy}, \alpha_{yt}$, and α_{tt} are parameters to be estimated. The price function $C(p)$ is the form that (15.30) would assume under constant returns to scale, when technical change is ignored.

Linear homogeneity in p can be imposed on the cost function defined by (15.30) and (15.31) via all the $N + 3$ parametric restrictions in (15.22). Then, as in the *TL* case, the key parameters allowing for nonhomotheticity (α_{iy}) and those capturing the bias of technical change (α_{it}) sum to zero. Thus, after taking into account the restrictions required by the homogeneity property, this nonlinear version of translog contains $N(N + 1)/2 + 2N + 3$ free parameters in all. This version has, consequently, the same degree of flexibility as the standard versions (both with exponential technical change).¹⁸

A unique feature of this nonlinear version of translog is that the cost-minimizing demand system is derived directly from Shephard's lemma (after dividing each demand equation by output), in terms of input-output ratios (and not in terms of input cost shares as in the standard version). The N -equation system of input-output ratios typified by the cost function defined by (15.30) and (15.31) can be written as

$$a_i(p, y, t)_{TLE} \equiv \alpha_0 \cdot \exp\{\alpha_t t + \alpha_{tt} t^2\} \cdot y^{(\alpha_y - 1 + 0.5\alpha_{yy} \ln y + \alpha_{yt} t)} \quad (15.32)$$

¹⁸Flexibility here is meant in the sense of Diewert. It is measured in terms of the number of free parameters in the model. See Diewert and Wales (1987).

$$\begin{aligned} & \cdot (\alpha_i + \sum_{j=1}^N \alpha_{ij} \ln p_j + \alpha_{iy} \ln y + \alpha_{it} t) / p_i \\ & \cdot \prod_{i=1}^N p_i \quad (\alpha_i + 0.5 \sum_{j=1}^N \alpha_{ij} \ln p_j + \alpha_{iy} \ln y + \alpha_{it} t), \quad \forall i \end{aligned}$$

The cost-minimizing share equations can be deduced from (15.30) and (15.32) using the relation: $S_i = p_i a_i / c$, where $c = C/y$. However, unlike the share system (15.20), the N -equation system of input-output ratios (15.32) embodies all the empirical implications of the cost function.

15.3.3 The Second Nonlinear Version of Translog (TLA)

With this second nonlinear version of the translog cost function (TLA) (which is can be interpreted as a modified form of the nonlinear version presented above) *stochastic specification* and *technical change* modeling are simultaneously altered, in comparison with the standard version *TL*. Its general formula allowing for nonhomotheticity and nonneutral technical change is given by

$$\begin{aligned} C(p, y, t)_{TLA} \equiv & C(p) \cdot \exp \{ \alpha_t t + \alpha_{tt} t^2 \} \cdot y^{(\alpha_y + 0.5 \alpha_{yy} \ln y + \alpha_{yt} t)} \quad (15.33) \\ & + \sum_{i=1}^N \psi_{iy} p_i \ln y + \sum_{i=1}^N \psi_{it} p_i t \end{aligned}$$

where the price function $C(p)$ is defined by (15.31), and be estimated.

Applying Euler's theorem on homogeneous functions, the cost function defined by (15.33) and (15.31) is linear homogeneous in p if and only if the following $N - 1$ parametric restrictions are imposed

$$\sum_{i=1}^N \alpha_i = 0, \quad \sum_{j=1}^N \alpha_{ij} = 0, \quad \forall i \quad (15.34)$$

Theorem 1. *The modified translog cost function defined by (15.31), (15.33) and (15.34) is a positive real valued cost function and linearly homogeneous in p . This means that for all real $\lambda > 0$, $y > 0$ and $p \gg 0_N$, we have:*

$$C(\lambda \cdot p, y, t)_{TLA} = \lambda \cdot C(p, y, t)_{TLA}, \quad \text{where } \lambda \cdot p \equiv (\lambda p_1, \lambda p_2, \dots, \lambda p_N). \quad (15.35)$$

Proof.

$$\begin{aligned}
 C(\lambda \cdot p, y, t)_{TLA} &= \exp(\alpha_t t + \alpha_{tt} t^2) \cdot y^{(\alpha_y + 0.5\alpha_{yy} \ln y + \alpha_{yt} t)} \cdot C(\lambda \cdot p) \\
 &\quad + \lambda \cdot \sum_{i=1}^N \psi_{iy} p_i \ln y + \lambda \cdot \sum_{i=1}^N \psi_{it} p_i t
 \end{aligned} \tag{15.36}$$

From (15.36), to demonstrate that $C(p, y, t)_{TLA}$ defined by (15.31), (15.33) and (15.34) is linearly homogeneous in p , we need only to demonstrate that $C(p)$ is homogeneous of degree one in p :

$$\begin{aligned}
 C(\lambda \cdot p) &= \alpha_0 \cdot \prod_{i=1}^N (\lambda \cdot p_i)^{(\alpha_i + 0.5 \sum_{j=1}^N \alpha_{ij} \ln(\lambda \cdot p_j))} \\
 &= \alpha_0 \prod_{i=1}^N (\lambda \cdot p_i)^{(\alpha_i + 0.5 \sum_{j=1}^N \alpha_{ij} \ln \lambda + 0.5 \sum_{j=1}^N \alpha_{ij} \ln p_j)} \\
 &= \alpha_0 \prod_{i=1}^N (\lambda \cdot p_i)^{(\alpha_i + 0.5 \sum_{j=1}^N \alpha_{ij} \ln p_j)} \\
 &= \prod_{i=1}^N \lambda^{(\alpha_i + 0.5 \sum_{j=1}^N \alpha_{ij} \ln p_j)} \cdot C(p) \\
 &= \exp\{\ln \lambda \cdot \sum_{i=1}^N (\alpha_i + 0.5 \sum_{j=1}^N \alpha_{ij} \ln p_j)\} \cdot C(p) \\
 &= \exp\{\ln \lambda \cdot (1 + 0.5 \sum_{i=1}^N \sum_{j=1}^N \alpha_{ij} \ln p_j)\} \cdot C(p) \\
 &= \lambda \cdot C(p)
 \end{aligned}$$

□

However, with this modified translog cost function, the linear homogeneity in p does not impose the adding-up restrictions on the parameters allowing for nonhomotheticity (ψ_{iy}) and nonneutral technical change (ψ_{it}). Those parameters

affect *independently* each input. For example, the key parameter (ψ_{it}) which captures the bias of technical change with respect to the i th input demand is input- i specific.

The cost function defined by (15.31) and (15.33) contains $N(N + 1)/2 + 2N + 5$ free parameters in all, after taking into account the restrictions in (15.34) required by the homogeneity property. Thus, this new version of translog with linear technical change (TLA) contains more parameters than the two other versions with exponential technical change (TL and TLE). Indeed, TLA is more flexible than TL and TLE.

The cost-minimizing input demand system typified by the TLA model is derived directly from (15.33) using Shephard's lemma. After dividing each demand equation by output, we get the following system of input-output ratios

$$a_i(p, y, t)_{TLA} \equiv \frac{C(p)}{p_i} \cdot \exp(\alpha_i t + \alpha_{tt} t^2) \cdot y^{(\alpha_y - 1 + 0.5\alpha_{yy} \ln y + \alpha_{yt} t)} \\ \cdot (\alpha_i + \sum_{j=1}^N \alpha_{ij} \ln p_j) + \psi_{iy} + \psi_{it} t y^{-1}, \quad \forall i \quad (15.37)$$

Assuming constant returns to scale, the unit cost formula is equivalent to

$$C(p, t)_{TLA} \equiv C(p) \cdot \exp\{\alpha_i t + \alpha_{tt} t^2\} + \sum_{i=1}^N \psi_{it} p_i t \quad (15.38)$$

However, the constant returns to scale property reduces the number of free parameters to $(N + 2)(N + 1)/2 + 1$.

We turn now to emphasize some existent differences between the TLE and TLA models tied to the technical change modeling aspects, in a purely analytic framework. For this purpose, the following notations are used

$$C(p) \equiv \alpha_0 \cdot \prod_{i=1}^N p_i^{(\alpha_i + 0.5 \sum_{j=1}^N \alpha_{ij} \ln p_j)} \quad (15.39)$$

$$S_i(p) \equiv \alpha_i + \sum_{j=1}^N \alpha_{ij} \ln p_j \quad (15.40)$$

$$\varphi(t) \equiv \exp(\alpha_i t + \alpha_{tt} t^2) \quad (15.41)$$

$$h(y, t) \equiv y^{(\alpha_y + 0.5\alpha_{yy} \ln y + \alpha_{yt} t)} \quad (15.42)$$

Hence, by using notations in (15.39)–(15.42), each of the two cost functions, TLE and TLA, together with their associated cost-minimizing share system and their associated input-output ratio system can be written in simplified forms

$$C(p, y, t)_{TLE} \equiv \varphi(t) \cdot h(y, t) \cdot C(p) \cdot \prod_{i=1}^N p_i (\alpha_{iy} \ln y) \cdot \exp(t \cdot \sum_{i=1}^N \alpha_{it} \ln p_i), \quad (15.43)$$

$$S_i(p, y, t)_{TLE} \equiv S_i(p) + \alpha_{iy} \ln y + \alpha_{it} t, \quad (15.44)$$

$$a_i(p, y, t)_{TLE} \equiv \varphi(t) \cdot \frac{h(y, t)}{y} \cdot \frac{C(p)}{p_i} \cdot (S_i(p) + \alpha_{iy} \ln y + \alpha_{it} t) \cdot \prod_{i=1}^N p_i (\alpha_{iy} \ln y) \cdot \exp(t \cdot \sum_{i=1}^N \alpha_{it} \ln p_i) \quad (15.45)$$

$$C(p, y, t)_{TLA} \equiv \varphi(t) \cdot h(y, t) \cdot C(p) + \sum_{i=1}^N \psi_{iy} p_i y + \sum_{i=1}^N \psi_{it} p_i t \quad (15.46)$$

$$S_i(p, y, t)_{TLA} \equiv [\varphi(t) \cdot h(y, t) \cdot S_i(p) \cdot C(p) + \psi_{iy} p_i y + \psi_{it} t p_i] / C(p, y, t)_{TLA} \quad (15.47)$$

$$a_i(p, y, t)_{TLA} \equiv \varphi(t) \cdot \frac{h(y, t)}{y} \cdot \frac{S_i(p) \cdot C(p)}{p_i} + \psi_{iy} + \psi_{it} t y^{-1} \quad (15.48)$$

The own and cross price elasticities of demand typified by the *TLE* and *TLA* models can be written as

$$\epsilon_{ij}^{TLE} \equiv \frac{\alpha_{ij} - \delta_{ij} S_i(p, y, t)_{TLE} + S_i(p, y, t)_{TLE} S_j(p, y, t)_{TLE}}{S_i(p, y, t)_{TLE}}, \quad \forall i, j \quad (15.49)$$

$$\epsilon_{ij}^{TLA} \equiv \varphi(t) \cdot h(y, t) \cdot \frac{\alpha_{ij} - \delta_{ij} S_i(p) + S_i(p) S_j(p)}{S_i(p, y, t)_{TLA}} \cdot \frac{C(p)}{C(p, y, t)_{TLA}}, \quad \forall i, j \quad (15.50)$$

Thus, the specification of the technical change (exponential or linear) seems to be decisive in the expression of the derived elasticities.

15.3.4 The Generalized Leontief (GL)

The generalized Leontief model, proposed by Diewert (1971), provides an alternative approximation of the cost function:

$$C(p, t)_{GL} \equiv \sum_{i=1}^N \sum_{j=1}^N \beta_{ij} (p_i p_j)^{1/2} + \sum_{i=1}^N \beta_{it} p_i t, \quad \beta_{ij} = \beta_{ji}, \quad \forall i, j \quad (15.51)$$

where constant returns to scale are assumed for brevity in exposition. Note that the *GL* cost function is linearly homogeneous in input prices by construction, while the symmetry of the second order parameters assures the symmetry of the Hessian matrix. If, in addition, all the β_{ij} for $i \neq j$ are constrained to be nonnegative, then the *GL* cost function will be concave. However, these nonnegativity constraints reduce the flexibility of this form by ruling out possibility of complementarity between inputs.

The cost-minimizing input-output ratios equations derived from (15.51) are of the form:

$$a_i(p, t)_{GL} = \sum_{j=1}^N \beta_{ij} (p_j / p_i)^{1/2} + \beta_{it}, \quad \forall i \quad (15.52)$$

which are linear in the parameters.

15.3.5 The Symmetric Generalized McFadden (MF)

Assuming constant returns to scale and nonneutral technical change, the Symmetric Generalized McFadden approximation of the cost function, proposed by McFadden (1978) and Diewert and Wales (1987), takes the form:

$$C(p, t)_{MF} \equiv g(p) + \sum_{i=1}^N \gamma_i p_i + \sum_{i=1}^N \gamma_{it} p_i t \quad (15.53)$$

with,

$$g(p) \equiv \frac{\sum_{i=1}^N \sum_{j=1}^N \delta_{ij} p_i p_j}{\sum_{i=1}^N \theta_i p_i}, \quad \delta_{ij} = \delta_{ji}, \quad \forall i, j \quad (15.54)$$

where γ_i , γ_{it} , and δ_{ij} are unknown parameters, and the $\theta_i > 0$ are predetermined parameters, for all i, j .

Following Diewert and Wales (1987) and Diewert and Wales (1995), the price function g given in Eq. (15.54) is a quadratic function in prices which is normalized by a linear form, to yield a function that is homogeneous of degree one in

prices. Thus, the *MF* cost function defined by Eqs. (15.53) and (15.54) is linearly homogeneous in input prices by construction, while the symmetry of the Hessian matrix requires the symmetry of the second order parameters.

Yet, the *MF* cost function retains the main attractive property, which is that the correct curvature conditions could be imposed through parameter restrictions without destroying the flexibility of this functional form (see Diewert and Wales 1987).

The cost-minimizing system of input-output ratios typified by the *MF* cost function defined by (15.53) and (15.54) is derived by means of Shephard's lemma as

$$a_i = \gamma_i + \frac{\sum_{j=1}^N \delta_{kj} p_j}{\tilde{p}} - 0.5 S_i^0 \cdot \frac{\sum_{k=1}^N \sum_{j=1}^N \delta_{kj} p_k p_j}{\tilde{p}} + \gamma_{it} t, \quad \forall i \quad (15.55)$$

where, the remaining non-homogeneous restriction is needed for identification purposes: $\sum_{i=1}^N \delta_{ij} = 0$.

15.4 Estimation and Empirical Results

The proposed parametric specifications described previously are applied to annual time series data for the three industrial branches of French manufacturing: intermediate goods (B04), equipment goods (B05) and consumer goods (B06). The three data sets contain information covering the period 1970–1989 on output level (y) together with information on prices and quantities for three inputs: capital (K), labor (L) and energy (E). All data are constructed using annual series taken from the volumes of national accounting of French National Institute of Economic and Statistical Information (INSEE).¹⁹ The extracted sample period run from 1970 to 1989, so the most important fluctuations of energy prices before Gulf War are included. The prices of the three inputs are scaled to unity in 1980, the approximate midpoint of the samples. For each branch the scaled data is obtained by dividing the original price and multiplying the original quantity of each input by its original price in 1980. Table 15.2 presents sample means and other descriptive statistics for the variables needed in the estimation of the different models.

For each of the three models, additive stochastic disturbance terms are appended to each equation in the corresponding regression system. The disturbances are assumed to posses classical statistical properties. For *TLE* and *TLA* the three input-output ratio equations are estimated as a seemingly unrelated regression (SUR) system since all the parameters of the cost function are recovered from the demand

¹⁹The method for calculating the user cost of capital services is exposed in detail in Baccar (1995b, 2003).

Table 15.2 Sample summary statistics: French manufacturing 1970–1989

Branch		Y	CU	p_K	p_L	p_E	a_K	a_L	a_E	S_K	S_L	S_E
Intermediate goods	Mean	229,881	0.944	1.271	1.070	0.912	0.150	0.592	0.217	0.194	0.605	0.201
	Std.	17,416	0.466	0.802	0.652	0.516	0.005	0.133	0.006	0.042	0.040	0.036
	Min.	189,446	0.348	0.402	0.249	0.237	0.140	0.409	0.206	0.127	0.538	0.150
	Max.	257,505	1.543	2.581	2.133	1.739	0.164	0.854	0.226	0.260	0.681	0.258
Equipment goods	Mean	205,760	0.948	1.262	1.093	0.964	0.098	0.837	0.042	0.125	0.835	0.040
	Std.	29,111	0.456	0.814	0.694	0.550	0.010	0.225	0.004	0.037	0.038	0.004
	Min.	141,406	0.377	0.387	0.246	0.272	0.086	0.539	0.036	0.071	0.772	0.034
	Max.	244,236	1.596	2.651	2.283	1.754	0.115	1.284	0.051	0.190	0.891	0.045
Consumer goods	Mean	147,219	0.856	1.277	1.106	0.958	0.097	0.717	0.054	0.137	0.808	0.055
	Std.	12,078	0.419	0.817	0.708	0.556	0.005	0.176	0.009	0.032	0.029	0.005
	Min.	115,010	0.340	0.399	0.251	0.255	0.090	0.496	0.041	0.084	0.764	0.045
	Max.	161,662	1.495	2.602	2.325	1.765	0.107	1.076	0.070	0.188	0.858	0.063

Y is the output quantity index, CU is the unit cost of production, p_i is the price index of input i ($i = K, L, E$ which stands for capital, labor and energy respectively), a_i is the i th input-output ratios, S_i is the i th input cost share

All input prices are normalized to unity for 1980

All input and output quantities are expressed in 1980 French francs

system. For *TL*, the log cost function and two corresponding cost share equations are estimated as a SUR system, one share equation is dropped prior to estimation. Iterated seemingly unrelated regressions are used to estimate the parameters for each system.²⁰ Because the covariance matrix was iterated, the estimates are invariant to which share equation dropped.

All of the three models have been estimated by imposing linear homogeneity and symmetry of the Hessian matrix on the cost function. The hypothesis of constant returns to scale and a nonneutral technical change are also imposed for the estimation while the first and second order autonomous terms on the time variable were deleted from all the model specifications (*TL*, *TLE*, *TLA*). Thus, in *TL*, under the maintained hypothesis of symmetry and separated linear homogeneity in p and y , the second order Taylor expansion is complete in input prices but incomplete in time (that is the exogenous indicator of “technical change”).

In this section, the cost-minimizing demand system is estimated in three versions: in terms of shares derived from the logarithm of the cost function for the log-quadratic model *TL*, (that is the usual version which associates a multiplicative error term with the estimated cost function and imposes an exponential technical change biases), and in terms of input-output ratios derived directly from the cost function for the exponential of the log-quadratic model *TLE* (that is the first nonlinear version with additive error term and exponential technical change biases) and for the modified translog *TLA* (that is the second nonlinear version with additive error term and linear technical change biases).

The first nonlinear form, denoted by ‘*TLE*’, associates an *additive* error term and an *exponential* technical change with the estimated cost function. It is from an analytical point of view strictly equivalent to the usual log-quadratic form, the difference laying in the stochastic specification. As in the *TL* form, the technical change parameters with respect to input prices sum to zero.²¹ This version has, consequently, the same degree of flexibility as the standard version *TL*.

The second nonlinear form, denoted by ‘*TLA*’, has an *additive* error term and a *linear* technical progress affecting *independently* each input. The interest of this modified translog is that the restrictions assuring homogeneity do not affect the technical change parameters. This specification is indeed more flexible than *TL* and *TLE* (both are with exponential technical progress).

The first interest of these two nonlinear versions of translog is that the derived input demand are expressed in terms of input-output ratios. So, it is not necessary in this case to join the cost equation with the system of regression, and more important, it avoids the problem of singularity of the cost-minimizing demand system involved by the estimation of the share equations system. The second interest, which is more important for our purpose, comes from the fact that they make us able to dissociate

²⁰The SYSNLYN procedure (option ITSUR) contained in SAS was used for the estimation of the different models.

²¹Imposing this kind of restrictions on the parameters of the input share equations, including those for the technical progress, is unavoidable for assuring linear homogeneity of the *TL* cost function.

the effect of stochastic specification (comparison between *TL* and *TLE*) from the pure effects of technical change modeling (comparison between *TLE* and *TLA*) on the reliability of estimates. Thus, they can be used as a vehicle to study potential biases resulting from the use of translog.

The performances of our three models, in approximating the true technology, are evaluated according to three principal criteria: (1) from a theoretical point of view, by testing if the regularity conditions validate the interpretation of the technology by means of a cost function; then, (2) from an economic point of view, by examining values, derivatives and temporal evolutions of price elasticities and Allen substitution elasticities, and finally (3) from a statistical point of view, by comparing the quality of fitting input shares, input-output ratios and unit cost.

15.4.1 Regularity Conditions

The estimated cost function should be well behaved, displaying consistency with theoretical regularity conditions (see Table 15.1) within the range of the observed data. Linear homogeneity in input prices essentially means that expressing prices in cents instead of Francs does not change actual behavior. The symmetry of the parameters guarantees the symmetry of the Hessian matrix. Monotonicity and concavity in input prices validate the assumption of cost minimizing-behavior.

In the remainder of the paper linear homogeneity in input prices together with symmetry of parameters across equations are imposed as maintained hypothesis in estimation for all of the cost systems. Thus I use economic theory in the well order to the largest possible extent *a priori*.²² However, imposing the appropriate curvature conditions (monotonicity and quasi-concavity) destroys the flexibility of the translog cost function. This loss of flexibility may, in turns, seriously biased the estimates.²³ I limit here myself to check the remaining properties at each point of the data.

The monotonicity property implies the nondecreasingness of the cost function in its input price arguments. Linear homogeneity and symmetry restrictions being imposed, I found that the three versions of the translog cost function are monotonic in the neighborhood covered by our set of the French manufacturing data. That is

²²The motivation for imposing the restrictions implied by symmetry and linear homogeneity in input prices are three. The first is the more pressing and it is that reported results and policy recommendations at least appear reasonable. The second motivation is to gain statistical efficiency in the estimation of parameters. Finally, imposition of these properties globally is very easy in practice since it involves simple linear restrictions on the parameters.

²³For instance, Diewert and Wales (1987) showed that the use of Jorgenson and Fraumeni (1981) procedure for imposing concavity on the translog cost function at all positive input price space (globally), not just in the historical sample on which the parameter estimates are based (locally), will seriously restrict the substitution possibilities allowed for by the technology and will generate a Cobb-Douglas representation of technology over certain areas.

Table 15.3 Concavity violations

Branch	<i>TL</i>	<i>TLE</i>	<i>TLA</i>
Intermediate goods	20	7	12
Equipment goods	9	4	6
Consumer goods	8	1	6

Concavity violations are sample points at which the estimated cost function is not concave. The sample size is 20

the fitted input cost shares derived from *TL* and the fitted input-output ratios derived from *TLE* and *TLA* are all positive for all three branches at all sample years.

The estimated cost function is concave in its input price arguments if its Hessian matrix is negative semi definite: thus, all principal minors of the Hessian matrix must alternate in sign starting with a negative value while diagonal elements must be all nonpositive. Because the cost function is linearly homogeneous in prices, the determinant of its Hessian will be zero. Therefore only the signs of the minors of dimension ($N - 1$) and smaller were tested.

Table 15.3 contains the number of sample points violating concavity. In average, this property is often violated by the three versions of translog in the neighborhood covered by our set of the French manufacturing data. Nevertheless, the results concerning this property are in all cases more satisfactory with the two nonlinear versions *TLE* and *TLA*.²⁴ The results obtained for *TL* are not shocking, since the estimated translog cost functions frequently fail to satisfy the concavity in prices (property that well-behaved cost function must possess), as has been shown in a number of empirical studies.²⁵

15.4.2 Economic Performances

The point that deserves examination, for a comparative analysis of the economic performances of the three versions of translog, is the coherence between the three types of estimations in terms of economic implications. This comparison is greatly facilitated by the evaluation of the elasticities of demand and of substitution derived from each of the three models. The analysis is done in three steps. We examine first their levels at the sample mean point, then their derivatives, and finally, for a deeper understanding, we focus our attention on their evolutions over the sample period.

²⁴In another experience, I found that concavity is satisfied by *GL* and *MF* throughout the historic period for equipment and consumer goods, but is often violated for intermediate goods.

²⁵See for example Berndt and Khaled (1979), Jorgenson and Fraumeni (1981), and Diewert and Wales (1987).

Table 15.4 Price elasticities and Allen partial elasticities of substitution for intermediate goods, French manufacturing, 1970–1989

Elasticity	<i>TL</i>	<i>TLE</i>	<i>TLA</i>
ϵ_{KK}	-0.068 ^a (0.010)	-0.194 ^a (0.021)	-0.105 ^a (0.015)
ϵ_{KL}	0.110 ^a (0.012)	0.214 ^a (0.027)	0.103 ^a (0.018)
ϵ_{KE}	-0.042 ^a (0.015)	-0.020 (0.014)	0.002 (0.013)
ϵ_{LK}	0.035 ^a (0.004)	0.066 ^a (0.008)	0.032 ^a (0.005)
ϵ_{LL}	-0.054 ^a (0.008)	-0.083 ^a (0.012)	-0.030 ^a (0.007)
ϵ_{LE}	0.019 ^a (0.007)	0.017 (0.019)	-0.002 (0.006)
ϵ_{EK}	-0.041 ^a (0.014)	-0.019 (0.014)	0.001 (0.013)
ϵ_{EL}	0.057 ^a (0.020)	0.053 (0.030)	-0.006 (0.023)
ϵ_{EE}	-0.016 (0.029)	-0.034 (0.027)	-0.005 (0.021)
σ_{KL}	0.182 ^a (0.020)	0.350 ^a (0.045)	0.224 ^a (0.036)
σ_{KE}	-0.211 ^a (0.074)	-0.103 (0.073)	0.009 (0.085)
σ_{LE}	0.094 ^a (0.034)	0.087 (0.064)	-0.013 (0.043)

Elasticities are evaluated at means of exogenous variables. Standard errors are in parenthesis. ϵ_{ii} ($i = K, L, E$) means the own price elasticity of demand for input i , ϵ_{ij} ($i, j = K, L, E$) means the elasticity of demand for input i with respect to price of input j , σ_{ij} means the Allen elasticity of substitution between inputs i and j

^aSignificant at the 95 % level

15.4.2.1 Elasticities at Sample Mean Point

Tables 15.4, 15.5 and 15.6 report the own and cross price elasticities of demand and partial elasticities of substitution derived from the three models (*TL*, *TLE*, *TLA*) and estimated for the three branches of French manufacturing (intermediate goods, equipment goods and consumer goods).

Using a homothetic, non-neutral technical change specification for the cost function, the reported own-price elasticities of demand evaluated at mean sample point are all negative. That is the necessary condition for the concavity of the cost

Table 15.5 Price elasticities and Allen partial elasticities of substitution for equipment goods, French manufacturing, 1970–1989

Elasticity	<i>TL</i>	<i>TLE</i>	<i>TLA</i>
ϵ_{KK}	-0.077 ^a (0.028)	-0.145 ^a (0.031)	-0.069 ^a (0.022)
ϵ_{KL}	0.076 ^a (0.025)	0.130 ^a (0.032)	0.050 ^a (0.018)
ϵ_{KE}	0.001 (0.007)	0.015 ^a (0.008)	0.019 ^a (0.008)
ϵ_{LK}	0.011 ^a (0.004)	0.018 ^a (0.004)	0.007 ^a (0.002)
ϵ_{LL}	-0.029 ^a (0.004)	-0.033 ^a (0.005)	-0.015 ^a (0.003)
ϵ_{LE}	0.018 ^a (0.002)	0.015 ^a (0.002)	0.008 ^a (0.002)
ϵ_{EK}	0.004 (0.023)	0.044 ^a (0.023)	0.058 ^a (0.026)
ϵ_{EL}	0.368 ^a (0.045)	0.325 ^a (0.053)	0.187 ^a (0.037)
ϵ_{EE}	-0.372 ^a (0.039)	-0.369 ^a (0.046)	-0.245 ^a (0.034)
σ_{KL}	0.091 ^a (0.029)	0.154 ^a (0.038)	0.097 ^a (0.034)
σ_{KE}	0.030 (0.180)	0.376 ^a (0.189)	0.757 ^a (0.339)
σ_{LE}	0.440 ^a (0.054)	0.386 ^a (0.063)	0.337 ^a (0.070)

Elasticities are evaluated at means of exogenous variables. Standard errors are in parenthesis. ϵ_{ii} ($i = K, L, E$) means the own price elasticity of demand for input i , ϵ_{ij} ($i, j = K, L, E$) means the elasticity of demand for input i with respect to price of input j , σ_{ij} means the Allen elasticity of substitution between inputs i and j

^aSignificant at the 95 % level

function at this point (which is not observable) is respected by the three models. Note that the curvature conditions are only partially respected over the sample itself (see Table 15.3).²⁶

²⁶From an economic viewpoint the own price elasticity of demand is of vital importance. It gives a measure of input's conservation proportionate to an increase in that input's price, when output remains constant. The greater the substitutability of other factors for the input in question, the greater the opportunity for resource conservation ($\varepsilon_{ii} = -\sum_{i \neq j} \varepsilon_{ij}$).

Table 15.6 Price elasticities and Allen partial elasticities of substitution for consumer goods, French manufacturing, 1970–1989

Elasticity	<i>TL</i>	<i>TLE</i>	<i>TLA</i>
ϵ_{KK}	-0.091 ^a (0.023)	-0.155 ^a (0.024)	-0.067 ^a (0.021)
ϵ_{KL}	0.126 ^a (0.024)	0.197 ^a (0.027)	0.094 ^a (0.019)
ϵ_{KE}	-0.035 ^a (0.014)	-0.042 ^a (0.015)	-0.027 ^a (0.015)
ϵ_{LK}	0.021 ^a (0.004)	0.032 ^a (0.004)	0.015 ^a (0.003)
ϵ_{LL}	-0.052 ^a (0.007)	-0.068 ^a (0.008)	-0.041 ^a (0.005)
ϵ_{LE}	0.031 ^a (0.005)	0.036 ^a (0.005)	0.026 ^a (0.003)
ϵ_{EK}	-0.087 ^a (0.035)	-0.100 ^a (0.036)	-0.065 ^a (0.034)
ϵ_{EL}	0.452 ^a (0.076)	0.538 ^a (0.079)	0.398 ^a (0.051)
ϵ_{EE}	-0.365 ^a (0.064)	-0.438 ^a (0.065)	-0.333 ^a (0.055)
σ_{KL}	0.156 ^a (0.030)	0.243 ^a (0.033)	0.168 ^a (0.033)
σ_{KE}	-0.634 ^a (0.257)	-0.756 ^a (0.274)	-0.764 ^a (0.239)
σ_{LE}	0.560 ^a (0.094)	0.661 ^a (0.097)	0.744 ^a (0.104)

Elasticities are evaluated at means of exogenous variables. Standard errors are in parenthesis. ϵ_{ii} ($i = K, L, E$) means the own price elasticity of demand for input i , ϵ_{ij} ($i, j = K, L, E$) means the elasticity of demand for input i with respect to price of input j , σ_{ij} means the Allen elasticity of substitution between inputs i and j

^aSignificant at the 95 % level

For equipment goods (Table 15.5) and consumer goods (Table 15.6), the elasticities evaluated using the estimated parameters of the two nonlinear versions (*TLE* and *TLA*) have the same sign as those evaluated with the standard version (*TL*). These elasticities are always statistically significant, an exception is the capital-energy elasticity which is noninformative. Substitutability between all pairs of inputs is systematically observed for equipment goods, while complementary between capital and energy emerges for consumer goods. Still, the elasticity of substitution between these two inputs (σ_{KE}) is little informative: significant with *TLA* only for the equipment goods (Table 15.5) and with *TL* and *TLE* only for consumer goods (Table 15.6).

However, for intermediate goods (Table 15.4), a branch which is a great consumer of energy (see Table 15.2), the elasticities obtained by means of the three functional forms (*TL*, *TLE*, *TLA*) reveal more clearly the sensitivity of the estimates to the alternative modeling aspects. Even though the price-elasticities estimated using *TL* are almost always statistically significant, the elasticities of demand of energy (ε_{KE} , ε_{LE} and ε_{EE}) estimated with the nonlinear versions *TLE* and *TLA* are, in general, not significant.

The relation between energy and the other factors of production seems to be ambiguous in this branch. We conclude to a weak separability between energy and the two other factors (capital and labor): σ_{KE} and σ_{LE} are statistically not significant. This result does not seem to be offensive if one takes into account the nature of the allocation of resources. Indeed, the share of energy in the total cost of production in the intermediate goods is neatly superior to the cost share of this factor in the two other branches.

Even though the elasticities estimated for the three data sets have same signs with *TL* and *TLE*, the values obtained with *TLE* are superior to those obtained with *TL* and *TLA*. Thus the structure of the estimated technology appears slightly more flexible when one employs the nonlinear version with an exponential technical progress (*TLE*). This phenomenon is particularly striking for the consumer goods. Moreover, capital is more affected than the other inputs.

On the other hand, the results obtained by means of *TLA* are, in some cases, in contradiction with the results obtained by means of *TL* concerning the signs of the estimated elasticities. More precisely, this concerns the elasticities of substitution capital-energy and labor-energy estimated for intermediate goods.

To summarize, the stochastic specification of the model (comparison between *TL* and *TLE*) has limited consequences on the sign of the elasticities. By contrast, the results seem to be more sensitive to the modeling of the technical progress (comparison between *TLE* and *TLA*). Moreover, combining these two effects (the stochastic specification effect and the technical change biases modeling effect: comparison between *TL* and *TLA*), leads to contradictions in the interpretation of the relations between energy and the other inputs in intermediate goods (signs of σ_{KE} and σ_{LE}).

15.4.2.2 Variations of Elasticities

When one limits the analysis to the levels of the elasticities evaluated at the sample mean point, the results stemming from the three approximations of the cost function seems robust to the different aspects of modeling, at least for equipment goods and consumer goods. But the results reveals an inconsistency if not a contradiction between the estimates obtained for the intermediate goods. In particular, only the elasticities derived from *TLE* and *TLA* conclude to the existence of certain separability between energy and the two other inputs in this branch.

The evaluation of elasticities involves the use of first and second order partial derivatives of the cost function. Thus they are unable to inform us on the evolution

Table 15.7 Derivatives of price elasticities for consumer goods, French manufacturing, 1980

Elasticity	TL	TLE	TLA
ϵ_{KE}	-0.042 ^a (0.016)	-0.050 ^a (0.017)	-0.033 ^a (0.017)
$\partial\epsilon_{KE}/\partial p_K$ ^b	0.078 ^a (0.014)	0.077 ^a (0.012)	0.082 ^a (0.019)
$\partial\epsilon_{KE}/\partial p_L$	-0.100 ^a (0.011)	-0.093 ^a (0.009)	-0.078 ^a (0.003)
$\partial\epsilon_{KE}/\partial p_E$	0.021 ^a (0.005)	0.016 ^a (0.005)	0.029 ^a (0.002)
ϵ_{LE}	0.037 ^a (0.005)	0.042 ^a (0.005)	0.036 ^a (0.004)
$\partial\epsilon_{LE}/\partial p_K$	-0.015 ^a (0.002)	-0.015 ^a (0.002)	-0.015 ^a (0.002)
$\partial\epsilon_{LE}/\partial p_L$	-0.016 ^a (0.003)	-0.013 ^a (0.004)	-0.067 ^a (0.001)
$\partial\epsilon_{LE}/\partial p_E$	0.031 ^a (0.003)	0.028 ^a (0.004)	0.048 ^a (0.003)
ϵ_{EE}	-0.415 ^a (0.055)	-0.476 ^a (0.057)	-0.434 ^a (0.047)
$\partial\epsilon_{EE}/\partial p_K$	0.094 ^a (0.017)	0.086 ^a (0.015)	0.125 ^a (0.015)
$\partial\epsilon_{EE}/\partial p_L$	0.149 ^a (0.048)	0.102 ^a (0.042)	0.802 ^a (0.022)
$\partial\epsilon_{EE}/\partial p_E$	-0.243 ^a (0.054)	-0.188 ^a (0.049)	-0.493 ^a (0.040)

^aSignificant at the 95 % level. Standard errors are reported in parenthesis

^b $\partial\epsilon_{KE}/\partial p_K$ means the partial derivative of ϵ_{KE} with respect to the price of capital (p_K), evaluated at the 1980 year. The same criterion is applied for other entries

of the technological relationships between inputs facing exogenous shocks on input prices. This analysis is made possible by the evaluation of the derivatives of the elasticities relatively to input prices. This involves the use of the third order derivatives of the cost function.

Table 15.7 contains levels and variations of the elasticities of demand with respect to the price of energy (ϵ_{KE} , ϵ_{LE} and ϵ_{EE}) for consumer goods.²⁷ The

²⁷For brevity, I only report results for consumer goods. Note that the curvature properties required by theory are less violated in this branch.

corresponding estimates and standards errors are computed at the 1980 year, the approximated midpoint of the sample.²⁸

The levels and the derivatives of these elasticities obtained with the three versions of translog are all statistically significant. The own-price elasticity of energy seems in all cases highly sensitive to the variations of the price of this input (with the three forms). This elasticity is particularly sensitive to the price of labor in the case of *TLA*. The derivatives of the capital-energy elasticity suggest that the complementarity between capital and energy becomes more intense after an increase in the price of labor. Conversely, these two inputs tend to be less complements and perhaps substitutable when their own prices are increased. However, the relation of substitution between labor and energy seems less sensitive to the exogenous shocks on input prices.

To sum up, the derivatives of the elasticities estimated with the three versions of translog suggest that the variations of the elasticities (given by the sign of the derivatives) are relatively robust to the alternative stochastic specification and technical change biases modeling of translog. However, the levels of the derivatives of the elasticities are larger when one makes use of the version with linear technical progress (*TLA*). Therefore, the demand of inputs appear more variable relatively to the prices when the stochastic specification of the model and the technical change modeling are simultaneously altered.

15.4.2.3 Temporal Evolutions of Elasticities

The sensitivity of estimates to the stochastic specification of the model and to the alternative technical change modeling is also handled by of representing in the same graphics of the three curves obtained for each elasticity by estimating the three versions of translog.

Figures 15.1, 15.2 and 15.3 show the evolutions over the sample period of the capital-energy, labor-energy and energy-energy price elasticities obtained with the three models for consumer goods.

A visual inspection indicates that the curves of the elasticities estimated with *TLE* and *TLA* have the same shape as those estimated with the usual log-quadratic version *TL*. However, The observed differences, in terms of level, between the elasticities obtained by the three models, decreases from 1980 onwards. In particular, I obtain the same value at the year 1980 for the elasticity of demand of labor with respect to the price of energy with *TL* and *TLA* (Fig. 15.2). This is also true concerning the own-price elasticity of energy (Fig. 15.3)

²⁸However, at the sample mean point, the elasticities are evaluated at means of exogenous variables which correspond to the average of the logarithm of input prices in the *TL* case. Thus the analysis of their derivatives with respect to input prices at this point does not allow a useful interpretation. We have arbitrarily chosen the 1980 point which corresponds to the reference year of the sample: All input prices are equal to one at this point.

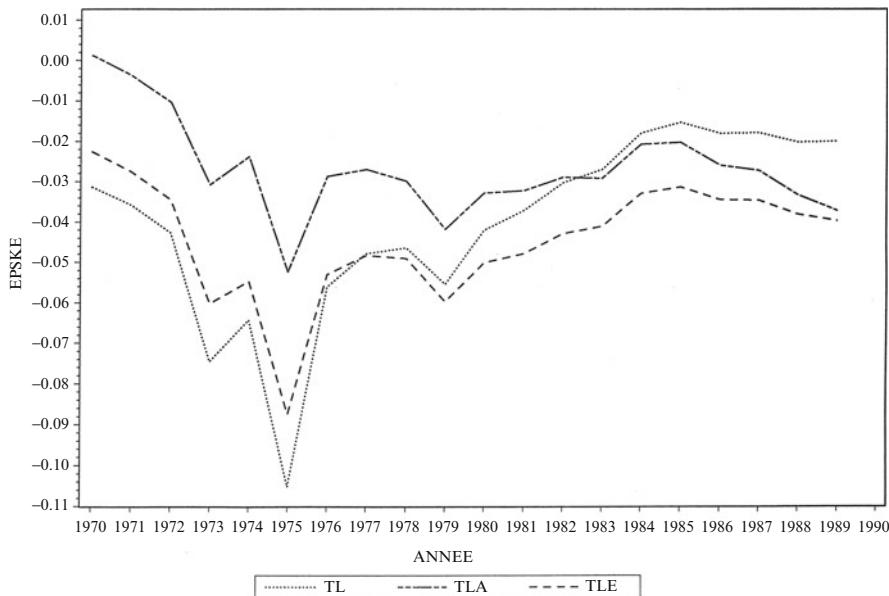


Fig. 15.1 Evolution of the elasticity of capital at the price of energy estimated with the three versions of translog for consumer goods

Anyway, these elasticities reveal some temporal variability concerning the extent to which inputs are substitute or complement and a neat variability for the own-price elasticity of energy. However, they keep the same sign throughout the sample period, at least for consumer goods detailed here.²⁹

The variations in the levels of these elasticities are more important for *TLA* (nonlinear with linear technical progress). It appears, again, that the estimated values for elasticities are more sensitive to the specification of the technical progress than to the stochastic specification of the model. In addition, for the three price elasticities considered here, which measure the impact of the energy price variations on the demand of each input, we note some reactions to the oil shocks. In particular, the sharp increase in the price of energy which began with the run-up of world petroleum prices in late 1973 caused an amplification in the reactions of input demands without modifying the nature of the technological relations between factors (the absolute values of elasticities become more important, but keep the same signs), at least for the consumer goods branch.

²⁹On the other hand, for the intermediate goods branch, variations in levels are accompanied with variations in signs.

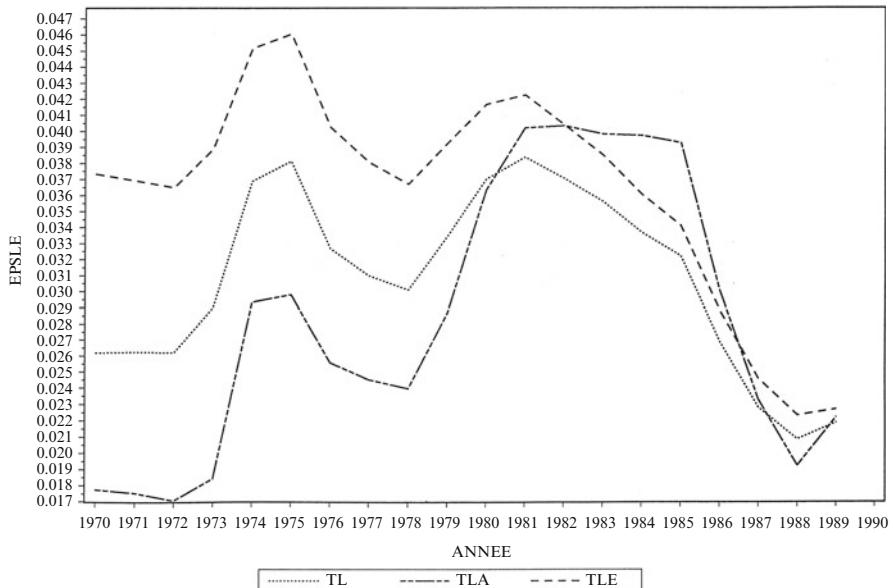


Fig. 15.2 Evolution of the elasticity of labor at the price of energy estimated with the three versions of translog for consumer goods

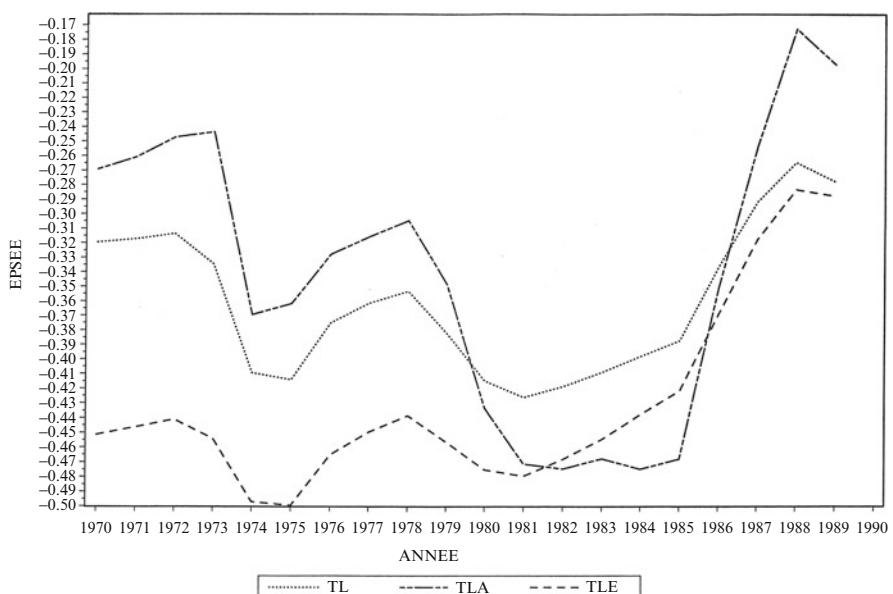


Fig. 15.3 Evolution of the own price elasticity of demand of energy estimated with the three versions of translog for consumer goods

15.4.3 Quality of Fit

After this detailed study of the sensitivity of the estimates to the version of translog used, first, in terms of mathematical performances (by testing the basic curvature conditions implied by theory), then in terms of economic performances (thorough a minutely analysis of price and Allen-Uzawa substitution elasticities), we address now the core of the problem: is the quality of fit improved when one uses one or the other nonlinear form?

A first idea about the quality of fit of the three models can be obtained by comparing the plots of observed and predicted values of unit cost, input-output ratios and shares against time. This visual exposition may provide a less ambiguous understanding of the differences between the estimated and the true values than would a summary statistic such as root mean square error which is sometimes employed in comparing predicted with actual values from time series data.

The observed cost shares of energy, energy-output ratio and unit cost and their predicted values, obtained by estimating the three specifications of translog for the consumer goods, are represented graphically in Fig. 15.4 through Fig. 15.6.

Inspection of the observed and predicted values for the cost share of energy (Fig. 15.4) indicates that the quality of fit provided by *TL* and *TLA* seems satisfactory. However, the results are less satisfactory with *TLE*: the share of energy is overestimated before 1979, while it is underestimated afterwards. This result is surprising since the quality of the fit poses no particular problem for the cost share, even when the estimated demand system was expressed in terms of input-output ratios (cases *TLA*, *GL* and *MF*). In fact, the exponential technical progress modeling seems incompatible with the additive stochastic specification.

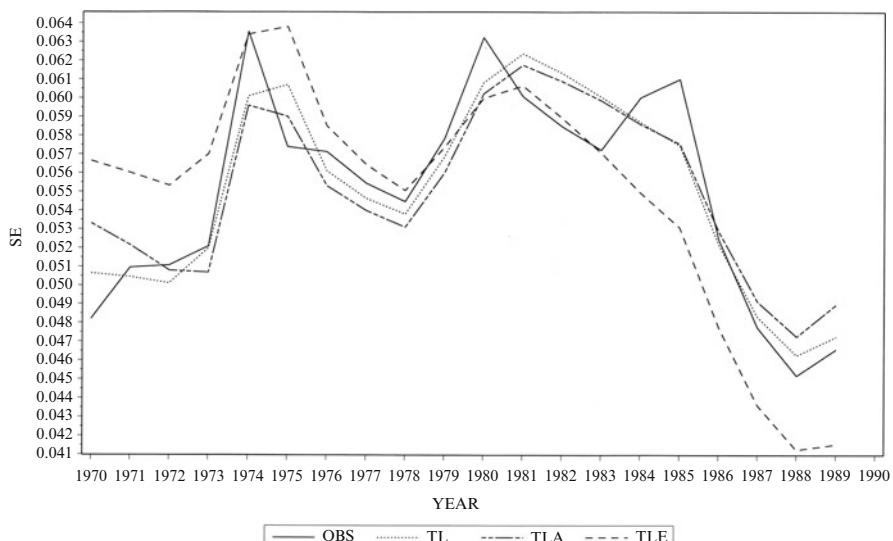


Fig. 15.4 Quality of adjustment of the cost share of energy obtained with the three versions of translog for consumer goods

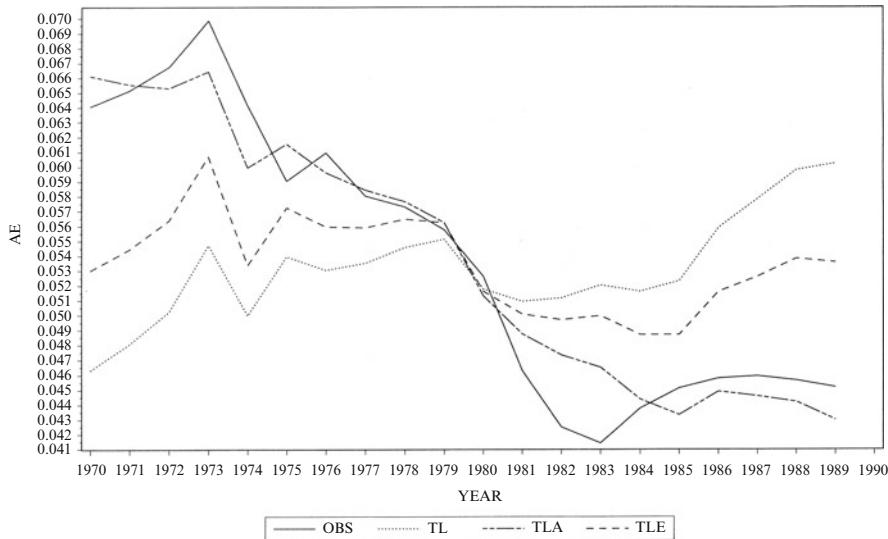


Fig. 15.5 Quality of adjustment of the energy-input ratios obtained with the three versions of translog for consumer goods

By contrast, the visual inspection of plots of observed and predicted energy-output ratio (Fig. 15.5) reveals that *TLA* outperform *TL* and *TLE*. In fact, the quality of fit of *TL* is mediocre while the improvement observed with *TLE* is not significant. The same conclusion comes out of Fig. 15.6 about the quality of fit of each of the three models.

We conclude that the little gain obtained by *TLE* in fitting input-output ratios is counterbalanced by a comparable loss in fitting input cost shares. In the three graphics (Figs. 15.4, 15.5, and 15.6), the predicted values obtained with *TLA* reproduce more closely the observed values. This result indicates that the use of an exponential representation of the technical change bias parameters is at the origin of the shortcomings of the log-quadratic model (*TL*).³⁰ This analysis supports the idea that the estimation of the cost-minimizing demand system in terms of input-output ratios while maintaining the exponential representation of the technical change biased does not resolve the problem.

15.5 Extensions on the Flexibility of the Log-Quadratic Model and Concluding Remarks

I complete now my search about the origin of the shortcomings of translog by reconsidering the parametrization used for the estimation of the model. In this section, I explore different degrees of flexibility of the usual log-quadratic form by

³⁰In fact, with *TL*, since the cost function is to be linear homogeneity in input prices, Euler theorem implies that: $\sum_{i=1}^N \alpha_{it} = 0$. Then the parameter α_{it} is not input i specific.

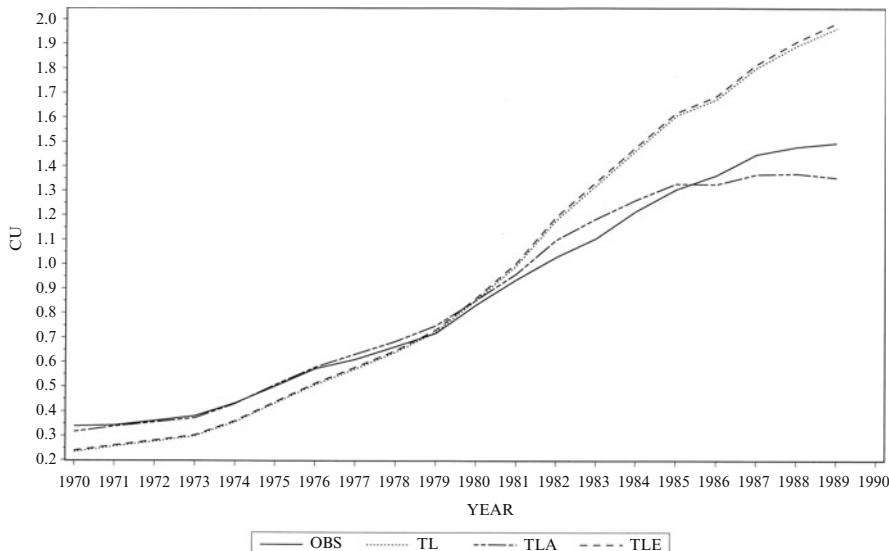


Fig. 15.6 Quality of adjustment of the unit cost obtained with the three versions of translog for consumer goods

varying the restrictions imposed on the technical change parameters. I concentrate in particular on the gain in precision obtained when the degree of flexibility relative to the time variable is increased.

Remember that the estimated translog cost function is “flexible” to the second order only with respect to input prices, while it is “semi-flexible” with respect to the time variable (that is the exogenous indicator of technical change) and “non-flexible” with respect to the level of output. The semi-flexibility in the time variable comes from the elimination of first order and second order autonomous technical change parameters from the model prior to estimation.³¹ The non-flexibility in the level of output comes from the assumption of constant returns to scale. Then the estimated unit cost function and its associated input cost shares are independent of output level.

Note that for the usual version of the translog cost function, only the parameters in interaction with the logarithms of input prices can be estimated from the system of input share equations. Thus, to estimate the autonomous parameters associated with the logarithm of output or time, it is necessary to incorporate the log cost function in the system of regression. It should be recalled that these parameters are key determinants in the computation of the rate of returns to scale and the overall

³¹See Diewert and Wales (1987, 1988) for more details about the notions of flexibility and semi-flexibility.

rate of productivity growth. Unfortunately, much of the empirical studies are based on the estimation of a system of input share equations.

There is another disadvantage in using the usual version of translog to estimate the parameters describing the technology. Since the cost function is to be homogeneous of degrees one in input prices, the adding-up condition implied by this property yields some supplementary restrictions on the parameters affected with the exogenous variables interacting with input prices. The parameters determining the bias of technical change and scale economies are included. Indeed, when the homogeneity constraints are imposed, the parameters associated to the time variable appearing in each input share equation must sum to zero. For this reason, it is impossible to estimate a technical change that affects independently each input when one uses a translog approximation. Thus, the translog model is, by construction, less flexible than other non-logarithmic models, such as generalized Leontief and symmetric generalized McFadden, which avoid this kind of restrictions on the parameters. This lack of flexibility apparently affects the quality of the estimates and can explain the lack of precision observed for these models in fitting the ‘input-output ratios and unit cost.³² Since the autonomous technical change parameters that appear in the cost function are not estimated, the effect of the time variable is not fully taken into account. This fact can explain the failure of the estimated model to fit unit cost and, consequently, the input-output ratios: passing from the share of an input in the total cost to its corresponding input-output ratio (in the *TL* case) or the converse (in the *GL* and *MF* cases), necessitates the use of the predicted cost recalculated from the estimated parameters. The question that I address now is the following: what is the effect of the inclusion of the parameters determining the autonomous technical change in the estimated cost function on the precision of translog in fitting input-output ratios and unit cost?

For this purpose, I consider three alternative parametrizations of the translog cost function with different degrees of flexibility obtained by imposing or relaxing restrictions about the parameters affecting the time variable. The three estimated models are: (1) a model with constant returns to scale and without technical change (*TL0*), (2) a model with constant returns to scale and technical change interacting with input prices only (*TL*), and (3) model (2) with relaxing the restriction imposed to the first order parameter of autonomous technical change (*TL1*).

I limit myself on comparing the observed values of input-output ratios and the unit cost to their predicted values obtained with the three models (*TL0*, *TL*, *TL1*). The corresponding curves are represented graphically in Figs. 15.7, 15.8 and 15.9, respectively.

A visual inspection of plots of observed and predicted unit cost and energy-output ratio against time suffices to be convinced that the quality of fit of the unit cost and of

³²The unit cost function is a linear combination of the endogenous variables of the systems of demand *GL* and *MF* expressed in terms of input-output ratios. By contrast, it is impossible to recalculate directly the unit cost or the input-output ratios from the system of input cost shares equations derived from *TL*.

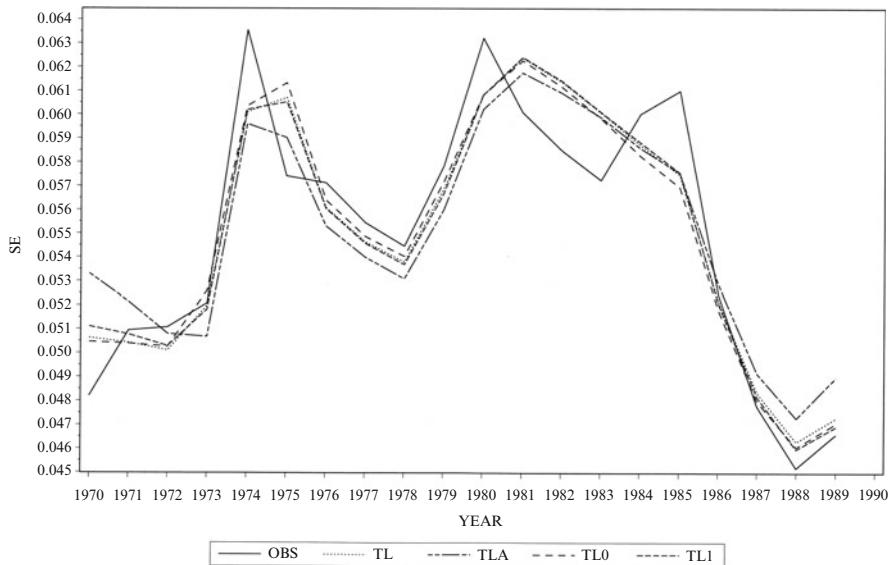


Fig. 15.7 Quality of adjustment of the cost share of energy obtained with different versions of translog for consumer goods

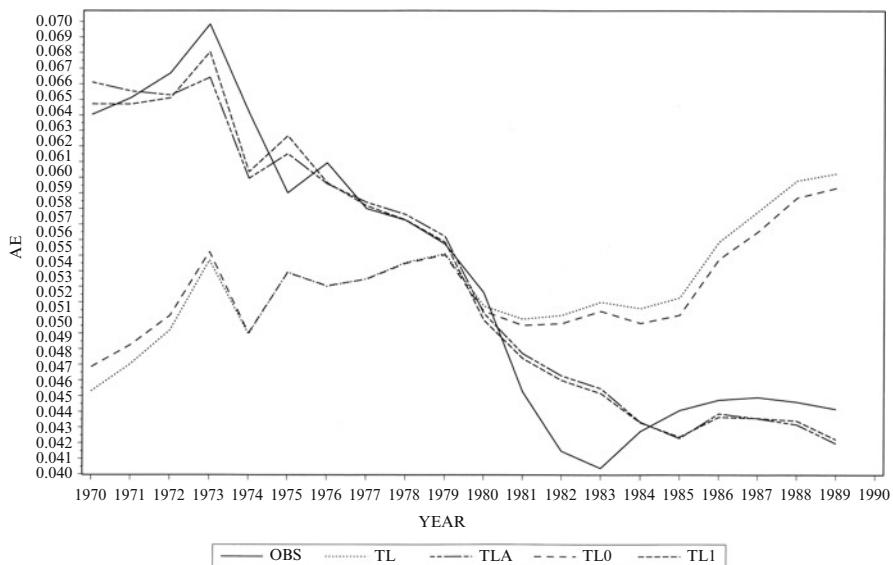


Fig. 15.8 Quality of adjustment of the energy-input ratios obtained with different versions of translog for consumer goods

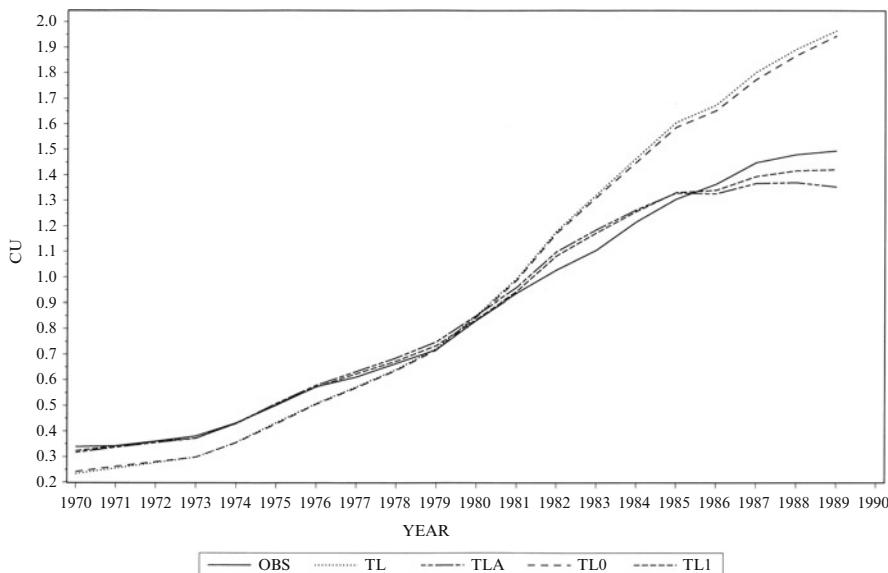


Fig. 15.9 Quality of adjustment of the unit cost obtained with different versions of translog for consumer goods

the input-output ratio of energy is neatly improved when the first order autonomous term is incorporated in the cost function. By contrast, the quality of fit is mediocre for the two other cases.³³

The main purpose of this study is to determine the sensitivity of empirical results and the reliability of fit to the alternative modeling aspects when estimations are based on a translog cost function: (1) alternative stochastic specification and technical change modeling by means of two nonlinear versions developed here; (2) the parametrization of the translog model used in the estimation (Figs. 15.10 and 15.11).

The results conclude that the stochastic specification of the model plays indeed a minor role in the quality of fit. However, the degree of flexibility of the estimated cost function, relative to the time variable (technical change), plays a significant role in the precision of fit of the model, in particular through the autonomous parameters. This confirms the results obtained with the two nonlinear versions.

³³Though the second order parameter of the autonomous technical change estimated for the three branches is always statistically significant, it has no effect on the precision of fit of translog when the first order autonomous parameter is estimated. A second experience relaxing the restriction on the returns to scale has been attempted. The gain in precision is modest.

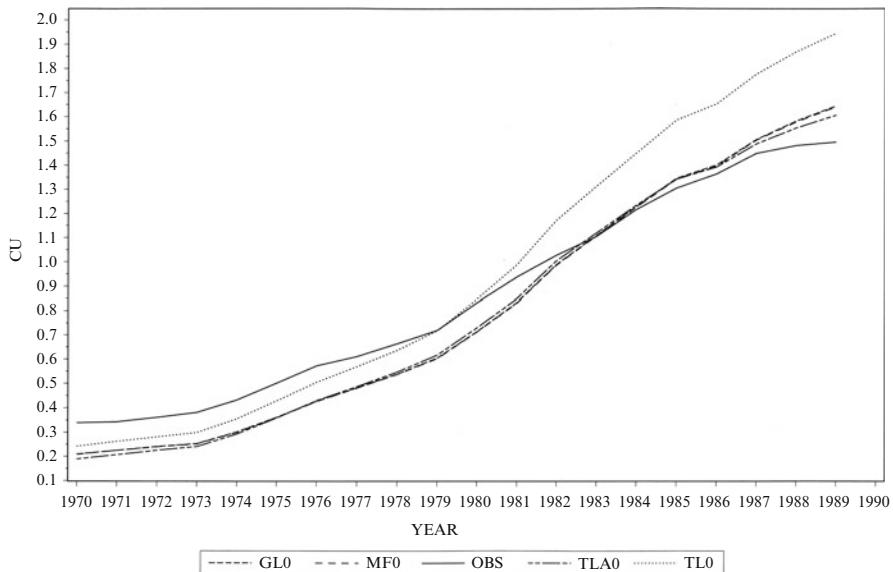


Fig. 15.10 Quality of adjustment of the unit cost obtained with different functional forms for consumer goods

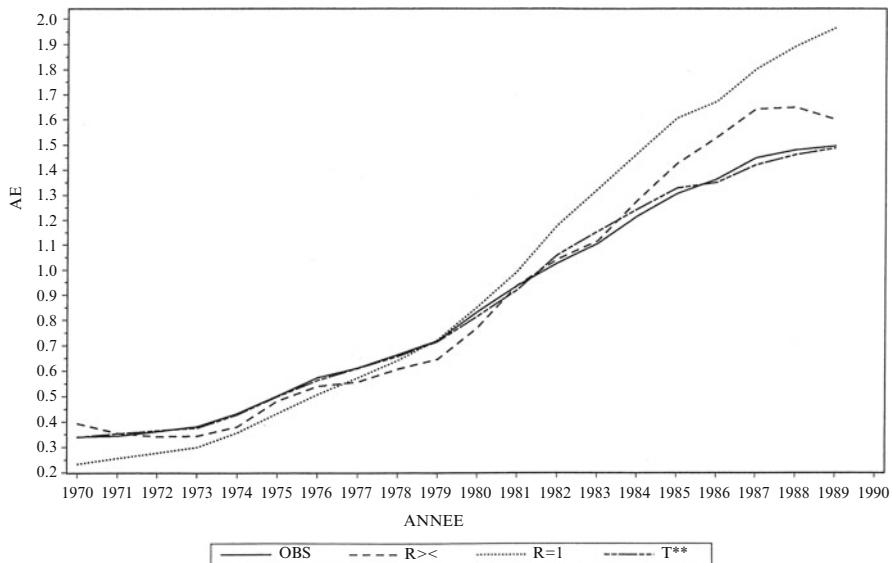


Fig. 15.11 Quality of adjustment of the unit cost obtained with different degrees of flexibility of the standard version of translog, when we relax the restrictions on technical change and returns to scale parameters, respectively

References

- Baccar S (1995a) Reliability of the translog cost function: some theory and an application to the demand of energy in french manufacturing. Centre de recherche en economie et statistique, CREST-INSEE working papers, vol 9547
- Baccar S (1995b) Coût d'usage du capital et fonction de Coût: la marge de liberté dans la modélisation laisse-t-elle des résultats robustes? Centre de Recherche en Economie et Statistique Seminar, CREST-INSEE
- Baccar S (2003) User cost of capital and cost function: does the margin in the modelling yields robust results? Computing in economics and finance 2003, 285, Society for Computational Economics.
- Baccar S (2006) Equilibrium specification and structure of technology: a factor substitution analysis in French industrial demand of energy. Computing in economics and finance 2006, 486, Society for Computational Economics.
- Barnett WA (1983a) New indices of money supply and the flexible laurent demand system. J Bus Econ Stat 1:7–23
- Barnett WA (1983b) Definitions of ‘second order approximation’ and of ‘flexible functional forms’. Econ Lett 12:31–35
- Barnett WA (1985) The minflex Laurent translog flexible functional form. J Econ 30:33–44
- Barnett WA, Jonas AB (1983) The Muntz-Szatz demand system. Econ Lett 11:337–342
- Barnett WA, Lee YW (1985) The global properties of the minflex Laurent, generalized leontief, and translog functional forms. Econometrica 53:1421–1437
- Barnett WA, Lee YW, Wolfe MD (1985) The three dimensional global properties of the minflex Laurent, generalized leontief, and translog functional forms. J Econ 30:3–31
- Berndt ER, Hesse D (1986) Measuring and assessing capacity utilization in the manufacturing sectors of nine OECD countries. Eur Econ Rev 30:961–989
- Berndt ER, Khaled MS (1979) Parametric productivity measurement and choice among flexible functional forms. J Polit Econ 87:1220–1245
- Brown RS, Christensen LR (1981) Estimating of elasticities of substitution in a model of partial static equilibrium : an application to U.S. agriculture, 1947 to 1974. In: Berndt ER, Fields BC (eds) Modeling and measuring natural resource substitution. MIT Press, Cambridge, MA, pp 219–268
- Caves DW, Christensen LR (1980) Global properties of flexible functional forms. Am Econ Rev 70:322–332
- Christensen LR, Jorgenson DW, Lau LJ (1971) Conjugate duality and the transcendental logarithmic production function. Econometrica 39:255–256
- Christensen LR, Jorgenson DW, Lau LJ (1973) Transcendental logarithmic production frontiers. Rev Econ Stat 55:28–45
- Diewert WE (1971) An application of the shephard duality theorem: a generalized Leontief production function. J Polit Econ 79:481–507
- Diewert WE (1974) Application of duality theory. In: Intriligator MD, Kendrick DA (eds) Frontier of quantitative economics, vol II. North-Holland, Amsterdam, pp 106–171
- Diewert WE (1982) Duality approaches to microeconomics theory. In: Arrow KG, Intriligator MD (eds) Handbook of mathematical economics, vol II. North-Holland, Amsterdam, pp 535–599; Edited by Intriligator MD, Kendrick DA. North-Holland, Amsterdam, pp 106–171
- Diewert WE, Wales TJ (1987) Flexible functional forms and global curvature conditions. Econometrica 55:43–68
- Diewert WE, Wales TJ Diewert WE, Wales TJ (1995) Flexible functional forms and tests of homogeneous separability. Journal of Econometrics 67: 259–302
- Diewert WE, Wales TJ (1988) A normalized quadratic semi-flexible functional form. J Econ 37:327–342
- Diewert WE, Avriel MA, Zang, I (1981) Nine kinds of quasi-concavity and concavity. J Econ Theory 25:397–420

- Diewert WE, Wales TJ (1995) Flexible functional forms and tests of homogeneous separability. *Journal of Econometrics* 67:259–302.
- Fuss MA, McFadden DL, Mundlak Y (1978) A survey of functional forms in the economic analysis of production. In : Fuss MA, McFadden DL, Mundlak Y (eds.) *Production economics: a dual approach to theory and applications*, vol I, North-Holland, Amsterdam, pp 219–268
- Gallant RA (1981) On the bias in flexible functional forms and an essentially unbiased form: the fourier functional form. *J Econ* 15:211–245
- Gallant AR, Golub EH (1984) Imposing curvature restrictions on flexible functional forms. *J Econ* 26:295–321
- Guilkey DK, Lovell CAK (1980) On the flexibility of the translog approximation. *Int Econ Rev* 21:137–147
- Guilkey DK, Lovell CAK, Sickles RC (1983) A comparison of the performances of three flexible functional forms. *Int Econ Rev* 24:591–616
- Jorgenson DW (1986) Econometric methods for modelling producer behaviour. In: Griliches Z, Intriligator MD (eds) *Handbook of econometrics*, vol III. North-Holland, Amsterdam, pp 1841–1915
- Jorgenson DW, Fraumeni B (1981) Relative prices and technical change. In: Berndt ER, Fields BC (eds) *Modeling and measuring natural resource substitution*. MIT Press, Cambridge, MA, pp 17–47
- Lau LJ (1974) Applications of duality theory: a comments. In: Intriligator MD, Kendrick DA (eds) *Frontiers of quantitative economics*, vol II. North-Holland, Amsterdam, pp 176–199
- Lau LJ (1986) Functional forms in econometric model building. In: Griliches Z, Intriligator MD (eds) *Handbook of econometrics*, vol III. North-Holland, Amsterdam, pp 1516–1566
- McFadden DL (1978) The general linear profit function. In: Fuss M, McFadden D (eds) *Production economics: a dual approach to theory and applications*, vol I. North-Holland, Amsterdam, pp 269–286
- Ohta M (1974) A note on the duality between production and cost function: rate of returns to scale and rate of technical progress. *Econ Stud Q* 25:63–65
- Rockafellar RT (1970) Convex analysis. Princeton University Press, Princeton
- Shephard RW (1953) Cost and production functions. Princeton University Press, Princeton
- Shephard RW (1970) Theory of cost and production functions. Princeton University Press, Princeton
- Uzawa H (1962) Production function with constant elasticities of substitution. *Rev Econ Stud* 29:291–299

Chapter 16

Bosman Ruling Implications on Player Productivity in the English Premier League

Mihailo Radoman

Abstract This paper examines the impact of policy changes on player productivity at the top level of European football, with a particular focus on the English Premier League. Contest theory motivates the prediction that post-Bosman entrants will be more productive and consequently have a higher probability of earning/retaining a first-team spot in top European leagues. To test these predictions, data was collected on all players that entered the English Premier League in four-year windows around the Bosman ruling. Nonparametric techniques, specifically Regression Discontinuity Design, were applied to test for sharp jumps in player productivity measures around the Bosman ruling; The results display evident discontinuity in player productivity measures, suggesting that post-Bosman entrants tend to be more productive than pre-Bosman entrants.

Keywords Productivity • Spillover hypothesis • Contest theory • Bosman ruling • Local average treatment effects • Regression discontinuity design

16.1 Introduction

The Bosman ruling in European football dates back to December 15, 1995. It served as the catalyst for major reform of the international transfer system in European sports, particularly in football. Among other changes, the ruling had a significant impact on regulations surrounding foreign player quotas. During the pre-Bosman period, clubs were restricted by the 3 + 2 rule in European competitions, meaning they can field three foreign players and two “assimilated” foreign players per match.¹ The judgement resulted in the removal of the “quota system” and European

¹Clubs were allowed to sign more than five foreign players, but faced these restrictions for each match-day. Similar restrictions were in place for domestic competitions as well.

M. Radoman (✉)
Department of Economics, Carleton University, 1125 Colonel By Drive,
Ottawa, ON, Canada K1S 5B6
e-mail: mihailoradoman@cmail.carleton.ca

clubs were allowed to field as many foreign nationals from EU countries as they desire.² A distinction can be made between the two periods surrounding the Bosman ruling, transitioning from a restrictive to a more open competitive environment in terms of player movement, akin to a move from autarky to free-trade. The post-Bosman period is characterized by a shift in power from clubs to players, which is reflected in the growth of player salaries in absolute terms and as a percentage of club revenues. Many European football experts have argued that these changes resulted in more uneven competitions across Europe. English press often suggest that the local youth development in the country has taken a step back to foreign talent, which is also reflected in the poor national team results on the international stage. This paper examines the impact of the Bosman ruling from a player's perspective, and focuses on the positive aspects of a more competitive environment that lead to significant growth in stature of major European football leagues. These leagues now attract the best talent from around the world and their popularity is becoming increasingly global. From an economic perspective, free-entry has created a more productive and efficient environment exemplified by increasing economic rent accruing to both players and clubs; not to mention the increase in appeal to the main driver of this industry, the sport fan or end consumer that indulges in the increased quality of this consumption good.

The post-Bosman period led to a greater presence of foreigners in all European leagues, which increased competition for places and the general talent level available for clubs to draw from, at least for higher quality leagues where clubs are not looking only at lower-cost alternatives to domestic journeymen. As a result of increased popularity, revenues from broadcasting rights increased significantly. Transfer values and salaries trended upwards as well, and one can argue that competition for spots in top European leagues intensified as a result. Whether or not these changes link directly to the Bosman ruling, their magnitude cannot be overlooked. For example, aggregate revenue for all English Premier League clubs has grown from £200 million in 1992/1993 to £1.3 billion in 2003/2004. Over the same period, wage costs have increased at a faster pace from £100 million (48 % of total club turnover) to £800 million (61 % of total club turnover). The increase in turnover has been universal across the major European competitions, while the wages/turnover ratio has increased as well in all competitions (particularly in Italy, moving from 57 % in 1995/1996 to 90 % in 2001/2002), with the exception of the German Bundesliga where this ratio remained relatively stable around 50 %. The growth in wages/revenues ratios became more apparent from 1995/1996 and on, especially for the English Premier League and Italian Serie A. The English Premier League distanced itself from other European competitions from 1995/1996 to

²Restriction remained for the number of foreign players originating from countries outside of the EU. However, in reality clubs would get around these rules as foreign players would get citizenship in a EU country that has more favourable immigration rules, which then allowed them to be considered as domestic players in any other EU leagues. For example, many South Americans easily attained Italian, Spanish, Portuguese citizenships.

2001/2002 in terms of revenue growth at an average rate of 22 %, from €534 million to €1.75 billion. The revenue from broadcasting rights increased significantly as well. The number of matches broadcasted in England increased from 60 per season from 1992–1997 to 138 when the new deal signed with Sky Sports in 1997.³

One of the main contributions of this paper is the uniquely created data set used to test the main theoretical predictions. Data is gathered for all players entering the English Premier League from 1992/1993⁴ to the 1999/2000 season. Their careers are carefully tracked for each season spent in a major European competition, and all of the relevant statistics are collected for each player. Non-parametric techniques are applied to test for treatment effects imposed by the Bosman ruling; Regression Discontinuity (RD) Designs are used to estimate the Bosman policy effects on player productivity measures. The RD approach has proven to be more than a program evaluation method in economics, providing a highly credible and transparent way of estimating program effects applicable to a wide variety of contexts. Examples of economic questions tackled by RD designs include: Labour supply effect of welfare, unemployment insurance and disability, education programs, median voter models, union effects on wages and employment, impact of free trade agreements on trade and product diversification, etc.. The Bosman ruling provides a natural cut-off point, not influenced or controlled by economic agents, that paves the way for the application of RD design. The results show a clear and sharp discontinuity in player productivity at the cut-off point. It is evident that post-Bosman market entrants exhibit increased productivity in footballing terms, outlined by a significant jump in their productivity measures. The results of the RD design are consistent with the theoretical predictions regarding a movement to a more homogeneous and productive player market. Theoretical motivation for the research question is found in relevant labour economics theory, including contest theory and positive-spillover hypothesis. Regardless of the theory, the underlying prediction of all models suggest that post-Bosman entrants will be more productive as a result of the removal of restrictive practises that were in place prior to the ruling.

Section 16.2 reviews the related literature and motivates the research question. Section 16.3 outlines the contents of the data set and provides basic summary statistics of the data. Section 16.4 presents the results of the regression discontinuity design, and Sect. 16.5 concludes the paper. The last two sections provide the tables and figures that are referenced in the text.

³All of the data above come from Deloitte & Touche, Annual Reviews of Football Finance.

⁴This coincides with the establishment of the Premier League in England. The newly established league had 22 teams in its inaugural season, after which the number of teams was reduced to 20.

16.2 Literature Review and Motivation

Globalization and opening up of labour markets has created many challenges and debates on the impact of immigration on the home country's economy. Positive aspects of this labour market research focus on the benefits gained by home economies from skilled-worker immigration, while negative arguments suggest that foreigners depress wages and reduce home-worker employment. Battu et al. (2003) presents the spillover hypothesis", whereby domestic workers permanently increase their human capital as a result of interaction with skilled co-workers from abroad that bring with them a new skill-set. European sports, particularly around the Bosman ruling, provide a natural testing environment for testing of many labour market questions about skilled-labour immigration. Alvarez et al. (2011) empirically test for the existence and strength of productivity spillovers from migrant to indigenous workers in European basketball, focusing on national team performance of host countries. Milanovic (2005) and Frick (2009) empirically address a similar question in European football, with mixed results on the impact of immigration on improving the competitiveness of national teams on the international stage. All of these studies test the implications of these labour market assumptions and predictions at the team level. It seems natural to question whether the immigrants are actually highly-skilled and superior to home-country players. Analyzing the labour market at the player level should shed light on some aspects of the spillover hypothesis, and the impact of structural changes on the competitive environment. The Bosman ruling opened up the EU countries and their sports leagues to a more skilled and wider talent pool. The question is whether this change resulted in a more competitive environment and higher quality/productivity of all players in a given EU league?

Competitions, in which rents are allocated as a function of the contestants efforts and ability in trying to win these rents, are a common economic phenomenon. Theoretical aspects of contests are generalizable and applicable to a wide range of activities. Examples of past studies examining strategic aspects of contests come from marketing, litigation, beauty contests, electoral competitions in politics, education filters, R&D contests, military conflict, sports, etc.. The competitive nature of sports provides a natural environment for the application of contest theory and testing of its predictions. Konrad (2007) provides a detailed survey of past work in contest theory, focusing on the strategic aspects of these games within a more general decision framework. Szymanski (2003) focuses his survey on the contest design issues specific to individual and team sports and outlines all of the work done in this field and many considerations that may change the nature of these contests. Franck and Cook (1995) showed that competition for limited number of positions in certain labour markets, including a spot on a National Basketball Association team roster, can lead to costly rent-seeking activities of competing players. Contest theory motivates the prediction that competition among a wider and more talented pool of players, such as in the post-Bosman period in Europe, will result in increased productivity among competing players, than competition in

a more restrictive environment, such as in the period preceding the Bosman ruling. The increased productivity might stem from higher average talent level and/or increased effort of post-Bosman entrants. As such, players entering the market prior to the ruling should be less productive, especially when directly competing for first-team spots with post-Bosman entrants.

16.3 Data Set and Summary

The empirical study is based on a uniquely collected data set consisting of all players that entered⁵ the English Premier League since its establishment in 1992/1993, ending with the 1999/2000 season. The concentration on new entrants aids in assessing players with similar characteristics outside of the changes brought upon by the Bosman ruling. The ruling itself occurred in December of 1995 (but effectively came into practice in January of 1996) and splits the data on entrants almost evenly on either side in terms of the seasons considered here. The career of each player that entered the English Premier League during this period is tracked and all publicly available and relevant statistics are captured for each season spent in a top European league.⁶ The data were collected using numerous internet sources, with the following sites serving as the backbone for the data generating process: <http://www.worldfootball.net/>, <http://www.transfermarket.co.uk/en>, <http://www.soccerbase.com/>. The information gathered includes: player's name, season of entrance ("year of entry") and each subsequent season spent in a top European league, player's source (whether he was brought up through the youth ranks, transferred in from a domestic club, or transferred in from a foreign club), the associated cost of acquiring his rights⁷ ("player cost"), each top-level club that a player appeared for in his career and relevant transfer fees paid for his rights, the ranking ("TeamPos") for each team that a player appeared for in each season, the share of foreign players in the top-level league for each relevant season ("ForShare"), player's position, age at entry (age of entry) and each subsequent season, nationality (UK, EU, and non-EU), appearances ("apps"), minutes played ("mpg"), goals scored, own-goals, yellow and red cards, international appearances ("intapps") and minutes played ("intmpg"), the player's transfer status for each season (whether the player stayed with the same club or transferred to another club

⁵Entering the league means making at least one first team appearance for any team in the English Premier League during the study period.

⁶Top European leagues consist of: English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A, and the French Ligue 1.

⁷For example, players brought up through the respective club's youth system are assigned a zero transfer value since the club does not have to bid for their rights in the transfer market.

Table 16.1 Player entries by Bosman and source

Player source	Pre-Bosman	Post-Bosman	Total
Domestic	130	110	240
Youth	148	152	300
Foreign	79	242	321
Total	357	504	861

for free or for a fee), and finally whether and for what reason⁸ a player exited the top-leagues in each season (most frequent reason for exiting was moving to a lower division team, whether it was by way of transfer or relegation).

Table 16.1 breaks down player entries by the Bosman ruling and player source. A total of 861 player entries are observed during the study period, of which 504 entered during the post-Bosman period and 357 entered during the pre-Bosman period. One statistic that particularly stands out is the number of foreign entrants in the post-Bosman period compared to the pre-Bosman period, which is more than three times higher (242 compared to 79). An increase was to be expected as the Bosman ruling removed restrictions surrounding player movement within the EU countries, but the magnitude of the increase in the first 4 years after the ruling certainly stands out. We observe that nearly half of all entrants are foreign sourced in the post-Bosman period compared to 25 % in the pre-Bosman period, which is predicted by the theoretical model as a result of easing the participation constraint for these players. The Bosman ruling did not impact the number of youth entrants in totality, but in relative terms as a percentage of total entrants it did have a negative effect of more than 10 %. Players sourced domestically suffered the most as the number of entrants decreased in total and relative numbers. The Bosman ruling removed certain trade barriers and significantly increased the talent pool that EU based clubs can draw from. Therefore, it is not surprising that there was a shift toward foreign players and that the number of entrants significantly increased in the post-Bosman period as the talent pool available to clubs grew in size and scope. The theoretical predictions and assumptions are consistent with these summary statistics and market information mentioned earlier.

Table 16.2 provides some basic summary statistics for certain variables in the data set, in total and broken down by pre and post-Bosman periods. Average appearances were higher by one for the post-Bosman entrants and average international appearances were higher by 0.5 as well for the same group. The average transfer cost for players entering on either side of the Bosman ruling was not materially different, and the highest observed transfer cost was around 40 million British pounds. The youngest player to enter the market during the study period was 16 and the oldest player still active at the top-level was 41. The average age of post-Bosman entrants was higher by just over 1 year compared to pre-Bosman entrants. The share of

⁸Players exiting the market due to injury are excluded from the final sample to improve the reliability of the statistical inference.

Table 16.2 Sample summary statistics

Bosman	Apps	IntApps	Age	TeamPos	Player cost	ForShare
<i>Pre-Bosman</i>	18.8	0.8	25.4	10.8	1.22 million	32.2
St. Deviation	11.9	2.1	4.4	6.1	3.04	2.5
Minimum	1	0	16	1	0	30.3
Maximum	42	13	39	22	40.5	37.1
<i>Post-Bosman</i>	19.7	1.4	26.4	10.4	1.95	49.1
St. Deviation	11.5	2.9	4.6	5.7	3.7	4.3
Minimum	1	0	16	1	0	37.1
Maximum	38	16	41	20	35.2	52.4
<i>Total</i>	19.3	1.1	26	10.6	1.66	42.2
St. Deviation	11.7	2.6	4.5	5.8	3.47	9.1
Minimum	1	0	16	1	0	30.3
Maximum	42	16	41	22	40.5	52.4

foreign players significantly increased in the post-Bosman period, approximately by 17 % at the mean level. The basic data analysis suggests that a further and more technical empirical approach is warranted in testing the effect of the Bosman ruling on the competitive structure of English and European football.

16.4 Regression Discontinuity (RD) Design

Estimating the existence and size of a discontinuous jump can be accomplished by comparing means in small bins of Z to the left and right of Z_0 or with a regression of various powers of Z , an indicator B for $Z > Z_0$, and interactions of all Z terms with B . A very basic method calculates the mean of each variable of interest for each bin on the time line selected, almost like a histogram. This rather simplistic visual approach paves the way for a more formal analysis below. Figures 16.1 and 16.2 present the means for two productivity proxies (appearances and minutes per game, respectively) for each of the four bins⁹ selected on either side of the Bosman ruling, with a quadratic fit line connecting the given bin means. The graphs suggest there is evidence of a jump in the conditional means for these two covariates around the Bosman ruling, particularly for the sub-sample of players that is nearest to the cut-off on both sides. Other productivity proxies, like international appearances, also display a discontinuity at the cut-off, while other variables examined (i.e. transfer costs) do not show signs of discontinuities.

The motivating theory suggests that there is a discontinuity in player productivity variables at the cut-off point. Players entering the market subsequent to the Bosman

⁹The results are consistent for other bins tested as well, but only the graphs for four bins are presented.

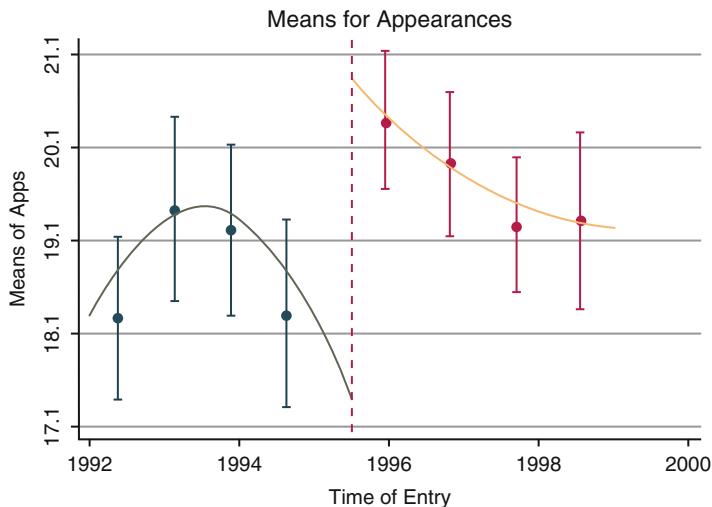


Fig. 16.1 Graph of bin means for Apps

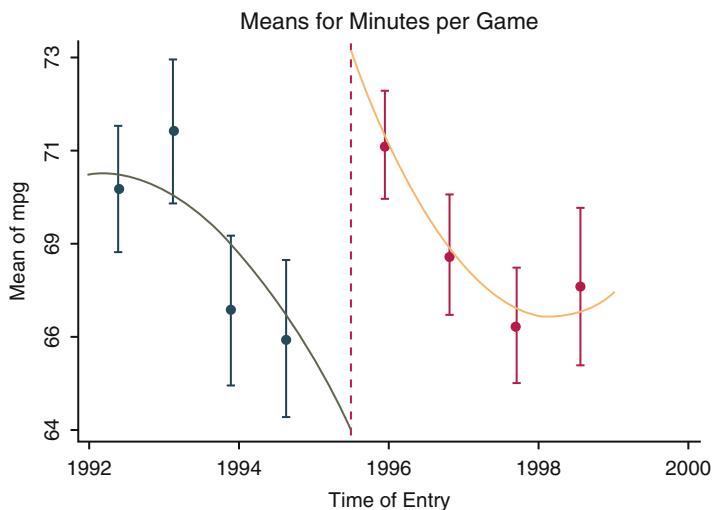


Fig. 16.2 Graph of bin means for MPG

ruling were subject to a different competitive and regulatory environment than players that entered the market ahead of the ruling. As such, the post-Bosman entrants are viewed as the treatment group; The institutional changes arising from the ruling are viewed as treatment effects to subjected post-ruling entrants (Table 16.3).

Table 16.3 Treatment effects

Treatment	Window	Outcome	Coefficient	Standard error	p-value
Bosman	1 year	mpg	4.68	1.41	0.001
Bosman	2 years	mpg	6.36	2.22	0.004
Bosman	3 years	mpg	7.04	1.79	0.000
Bosman	All years	mpg	6.38	1.54	0.000
Bosman	1 year	apps	2.08	0.76	0.006
Bosman	2 years	apps	2.76	1.19	0.021
Bosman	3 years	apps	2.64	0.96	0.006
Bosman	All years	apps	2.19	0.83	0.008
Bosman	1 year	intapps	0.45	0.17	0.009
Bosman	2 years	intapps	0.74	0.27	0.006
Bosman	3 years	intapps	0.33	0.22	0.134
Bosman	All years	intapps	0.13	0.19	0.505
Bosman	1 year	intmpg	5.32	2.06	0.003
Bosman	2 years	intmpg	9.68	3.27	0.010
Bosman	3 years	intmpg	4.24	2.62	0.106
Bosman	All years	intmpg	1.41	2.27	0.535

A proper RD design is characterized by a set of assumptions that are relevant in the context of this analysis:

- Two groups: Players entering before and after the institutional change represented by a dummy A
- Two periods of time around the cut-off point represented by a dummy variable T
- Players cannot precisely manipulate the assignment variable to influence whether they receive the treatment or not
- Treatment indicator defined by the dummy $B = AxT$
- Forcing and control variables that are continuous around the cut-off point

The identification of the policy impact on player productivity can be estimated by the following reduced form specification:

$$Y_i = \alpha + \beta B_i + f(X) + \varepsilon_i \quad (16.1)$$

where $B = 0$ for players entering the top tier of English football prior to the Bosman ruling, and $B = 1$ for players entering subsequent to the ruling. The assumption of continuity of the baseline covariate effect of a player's transfer value, X, around the cut-off is estimated by $f(X)$. Optimal bandwidth is obtained a la Imbens and Kalyanaraman (2009) method that minimizes the squared bias plus variance. The estimation is performed at four bandwidth levels around the cut-off, specifically at year 1, year 2, year 3, and year 4 on either side of the Bosman ruling. The different bandwidths serve as robustness checks in the estimation of the potential jump in the treatment outcomes around the Bosman ruling. Other control variables (i.e. the

relative league strength of the team employing the player) are used to test if the Bosman effect is appropriately estimated. The treatment effect is identified as:

$$\beta = \frac{\lim_{\varepsilon \downarrow 0} E[Y | B = \varepsilon] - \lim_{\varepsilon \uparrow 0} [Y | B = \varepsilon]}{\lim_{\varepsilon \downarrow 0} p(Z) - \lim_{\varepsilon \uparrow 0} p(Z)} \quad (16.2)$$

The treatment parameter is estimated using local polynomial regressions on either side of the cut-off point:

$$\min_{\beta} E[Y_i - \alpha - \beta(Z_i)]^2 K\left(\frac{Z_i - Z_0}{h}\right) \quad (16.3)$$

where $K(\frac{Z_i - Z_0}{h})$ is a triangle Kernel function that gives more weight to the observations closer to the cut-off, and h is the bandwidth level.

The results are presented in Table 16.2. The Bosman ruling generated positive productivity treatment effects as predicted by the theoretical model. Players entering the English Premier league in the post-Bosman period are playing approximately 6–7 more minutes per game and making 2–3 more appearances per season than the pre-Bosman control group.¹¹ All of the coefficients are statistically significant at the 5% level, with all but one significant at the 1% level. Variability is reduced for larger windows as expected, without visible losses to precision. Controlling for other covariates, like player age at entry and his respective team's ranking in the league, does not impact the results for the tested outcomes. In addition, the continuity assumption is satisfied for players' transfer costs (X) around the cut-off, as there is no evidence of a jump or discontinuity at that point. The table also presents results for international appearances and minutes per game, which display statistically significant results for 1 and 2 year windows at the 1% level, whereas the results for 3 and 4 year windows are insignificant at any reasonable level. This is not that surprising considering only a relatively small portion of players examined made international appearances during the study period. Nonetheless, the significant results for 1 and 2 year windows imply positive treatment effects in terms of player productivity even at the international level.

A graphical illustration of the results for minutes per game and appearances is presented in Figs. 16.3 and 16.4. The treatment effect at the cut-off is evident in all figures, and this result is robust to different bandwidth selections around the Bosman ruling. Figure 16.3 outlines the jump in minutes per game played by post-Bosman entrants versus pre-Bosman entrants, while Fig. 16.4 displays the discontinuous jump in appearances per season by players entering after the ruling.

¹⁰If the impact of treatment is immediate we have a sharp RD design (discrete jump at cut-off) and denominator becomes 1. If the effect is fuzzy around the cut-off, the denominator is in the [0, 1] interval.

¹¹The optimal bandwidth for mpg and apps was 2.12 and 2.46, respectively.

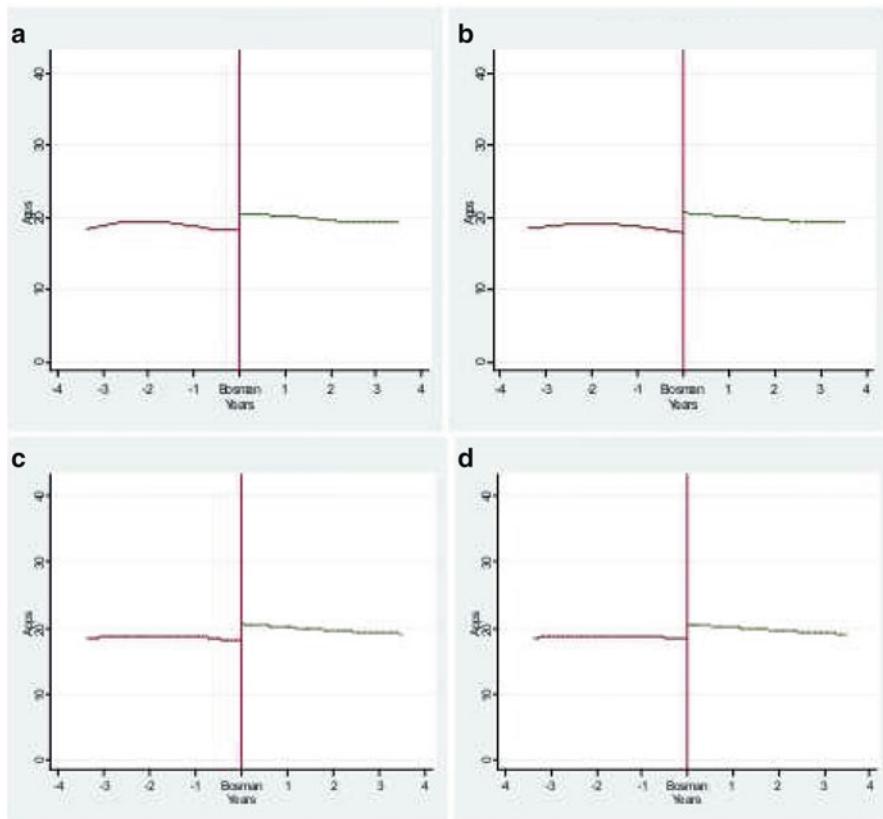


Fig. 16.3 (a) Apps for 1 year bandwidth, (b) Apps for 2 year bandwidth, (c) Apps for 3 year bandwidth, (d) Apps for 4 year bandwidth

The reduction (or removal in some cases) of barriers to free movement of labour increases the overall and average talent and effort levels of marginal entrants in the English Premier league.

16.5 Conclusion

The Bosman ruling has received much attention from many different fields, including academia, due to its importance in re-shaping the competitive structure of European sports. This paper accounts for the effects of the Bosman ruling as a structural shift that alters the nature of the competition. The removal of entry barriers for foreign players enhanced the pool of talent available to clubs to draw from, which intensified the competition in the English Premier League. As a result,

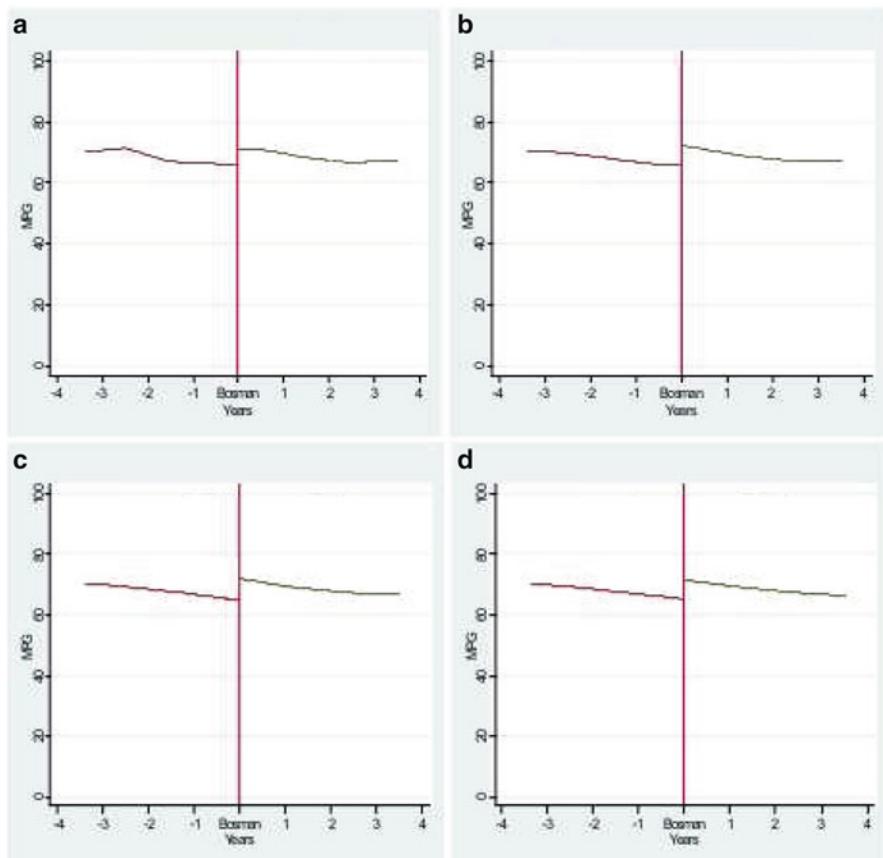


Fig. 16.4 (a) MPG for 1 year bandwidth, (b) MPG for 2 year bandwidth, (c) MPG for 3 year bandwidth, (d) MPG for 4 year bandwidth

the productivity levels necessary to survive at the top level increased as well, which translates into higher quality of average players in the league in the post-Bosman period. A uniquely collected data set is applied to test the research questions postulated. The results of the RD design non-parametric approach clearly indicate that there is a sharp jump in the two main productivity measures, appearances per season and minutes per game.

It is important to note that this analysis relies on a data set composed of only new entrants into the English Premier League, and cannot assess the impact of the Bosman ruling on all other incumbents in the league. Another limitation of the data is the lack of publicly available information on player salaries and other sources of income. An interesting extension to this paper could be to examine the impact of the Bosman ruling on a sub-group of players based on their source of entry. For example, the impact of the ruling on club owned youth academies

and the career success of the players they produce warrants additional research. Assessing the strength of youth academies and domestically sourced players versus foreign sourced players can address some of the public debates of football experts regarding suggested protectionist policies in European football, particularly in England. In addition, performing a similar analysis for other European competitions can be fruitful to assess whether there is consistency in the results across the top EU leagues.

References

- Alvarez J, Forrest D, Sanz I, Tena JD (2011) Impact of importing foreign talent on performance levels of local co-workers. *Labour Econ* 18:287–296
- Battu H, Belfield CR, Sloane PJ (2003) Human capital spill-overs within the workplace. *Oxf Bull Econ Stat* 65(5):575–594
- Franck H, Cook J (1995) The winner-take-all society. Free Press, New York
- Frick B (2009) Globalization and factor mobility: the impact of the Bosman Ruling on player migration in professional soccer. *J Sports Econ* 10:88–106
- Imbens G, Kalyanaraman K (2009) Optimal bandwidth choice for the regression discontinuity estimator. NBER WP 14726
- Konrad K (2007) Strategy in contests - an introduction. Social Science Research Center, Berlin
- Milanovic B (2005) Globalization and goals: does soccer show the way? *Rev Int Polit Econ* 12(5):829–850
- Szymanski S (2003) The economic design of sporting contests. *J Econ Lit* 41(4):1137–1187

Chapter 17

Productivity Measurement, Model Averaging, and World Trends in Growth and Inequality

Robin C. Sickles, Jiaqi Hao, and Chenjun Shang

Abstract Our paper provides new methods to robustify productivity growth measurement by utilizing various economic theories explaining economic growth and productivity and the econometric model generated by that particular theory. We utilize the World Productivity Database from the UNIDO to analyze productivity during the period 1960–2010 for OECD countries. We focus on three competing models from the stochastic frontier literature, Cornwell et al. (*J Econ* 46(1): 185–200, 1990), Kumbhakar (*J Econ* 46(1):201–211, 1990) and Battese, Coelli (*J Prod Anal* 3(1–2):153–169, 1992) to estimate productivity growth and its decomposition into technical change and efficiency change and utilize methods due to Hansen (2007) to construct optimal weights in order to model average the results from these three approaches.

Keywords Productivity • Panel data • Stochastic frontiers • Time varying heterogeneity • Model averaging • United Nations Industrial Development Organization

*This Plenary Session presentation was prepared for the North American Productivity Workshop VIII, hosted by Carleton University, Ottawa, Canada, June 4–June 7, 2014 and is based in part on Sickles, Hao, and Shang (Oxford Handbook on Panel Data, 2015), Shang (2015), Sickles et al. (2014). The authors would especially like to thank Marcel Viola for his extensive and careful editing of our contribution. The usual caveat applies.

R.C. Sickles (✉) • C. Shang
Department of Economics, Rice University, Houston, Texas, USA
e-mail: rsickles@rice.edu; Chenjun.Shang@rice.edu

J. Hao
Credit Scorecard and Portfolio Management, ATB Financial,
Edmonton, AB, Canada
e-mail: jhao@atb.com

17.1 Introduction

Proper measurement of nations' productivity growth is essential to understand current and future trends in world income levels, growth in per/capita income, political stability, and international trade flows. In measuring such important economic statistics is it also essential that a method that is robust to misspecification error is used. This talk addresses the robustification of productivity growth measurement by utilizing various economic theories explaining economic growth and productivity and the econometric model generated by that particular theory. We start from a realistic assumption that all models are misspecified in one way or another. Just as the famous quote by Box, "essentially, all models are wrong, but some are useful", carefully designed procedures to approximate the underlying DGP based on all collected information are needed. We address the heterogeneity problem by grouping countries according to their geographical, cultural and development characteristics as well as by the use of various panel data techniques. We utilize the World Productivity Database from the UNIDO to analyze productivity during the period 1960–2010. We consolidate the empirical findings from a number of statistical treatments consistent with the various economic models of economic growth and productivity. We discuss methodologies for averaging these various empirical findings. We also construct consensus estimates of world productivity TFP growth and find that, compared to efficiency catch-up, innovation plays a much more important factor in generating TFP growth.

17.2 Traditional Explanations for Sources of Economic Growth

The primary sources of economic growth and development are centered on two basic explanations: factor-accumulation and productivity-growth.

Rapid economic growth in East Asia in the 1970's and 1980's were thought by Kim and Lau (1994), Young (1992, 1995) and Krugman (1994) to be largely explained by the mobilization of resources. An alternative explanation to the neoclassical hypothesis explains economic growth in terms of intensive and extensive utilization of input factors as well as governmental industrial policies and liberalization policies. The sources of economic growth can be derived by explicitly introducing the role of catch-up due to an increase in productivity efficiency (Hultberg et al. 1999, 2004).

Introducing the role of efficiency in production means introducing some form of frontier production process, i.e., stochastic frontier production. Total Factor Productivity (TFP) growth is often decomposed into technological (technical innovation) change and technical efficiency change. Modifications of the neoclassical model can be found in the new growth theory. Endogenous growth models were developed to weaken the strong neoclassical assumption that long-run productivity growth

could only be explained by an exogenously driven change in technology. Sources of productivity differences in post WWII industrialized countries can be explained by neoclassical growth models that incorporate knowledge spillovers, technological diffusion, and convergence to a best practice production process (Smolny 2000).

17.2.1 Classical Residual Based Partial and Total Factor Productivity Measurement

Productivity historically has been specified as the ratio of some function of outputs (Y_i) to some function of inputs (X_i), which may be further adjusted by accounting for changing output and input mix. For a single such total factor productivity (TFP) output is often written as:

$$TFP = \frac{Y}{\sum a_i X_i}. \quad (17.1)$$

The a_i weights can be assigned as *arithmetic weighted averages* (Kendrick 1961) wherein the weights are typically based on expenditure shares, or as *geometric weighted averages* (Solow 1957). As growth in TFP is usually of primary concern, geometric averages have usually been used and this leads to the Solow measure, which is adopted by most central governments and statistical agencies and is based on the Cobb-Douglas production function with constant returns to scale, $Y = AX_L^\alpha X_K^{1-\alpha}$:

$$TFP = \frac{Y}{X_L^\alpha X_K^{1-\alpha}}. \quad (17.2)$$

Assuming cost minimization, the parameter α is the expenditure share of labor and a measure of TFP growth is the simple time derivative of TFP :

$$T\dot{F}P = \frac{dY}{Y} - \left[\alpha \frac{dX_L}{X_L} + (1 - \alpha) \frac{dX_K}{X_K} \right] \quad (17.3)$$

and thus a total factor productivity index is simply the difference between the log of the output index and the log of the input index. Growth in the index is thus the first difference over a time period in the differences of the log output aggregator and the log input aggregator (Jorgenson and Griliches 1972). Of course index numbers themselves have a long standing literature that has been surveyed by a number of scholars, based on part on the pioneering work of Fisher (1927) who formulated a number of desirable properties for index numbers. One such survey can be found in Good et al. (1997).

17.2.2 *Modifications of the Neoclassical Model: The New Growth Theory*

Endogenous growth models were developed to weaken the strong neoclassical assumption that long-run productivity growth could only be explained by an exogenously driven change in technology. An alternative interpretation to the endogenous growth literature is that it was a response to the simplistic view that the benefits of technical change (aka ‘manna from heaven’, Scherer 1971) were determined ‘outside the system.’ However, technological change as result of economic factors was discussed in Griliches’ 1957 Ph.D. dissertation and his concurrent article (Griliches 1957), wherein he pointed out that hybrid corn seed penetration followed a logistic distribution. The diffusion of innovations and the technological change it engenders has much in common with the penetration of seeds varieties in agricultural production. It is thus no surprise that numerous instances of such patterns were found by many researchers, including a productivity pioneer in his own right, Edwin (Mansfield 1961). Mansfield’s treatment of technological change and the rate of imitation was in its own right equally prescient. The classic model put forth by Romer (1986), which began the “new growth theory,” allowed for non-diminishing returns to capital due to external effects. For example, research and development by a firm could spill over and affect the stock of knowledge available to all firms. In the simple Romer model firms face constant returns to scale to all private inputs. The level of technology A can vary depending on the stock of some privately provided input R (such as knowledge) and the production function is formulated as

$$Y = A(R)f(K, L, R). \quad (17.4)$$

In the “new” growth theory the production frontier is shifted by factors that are endogenous, such as “learning-by-doing” (Arrow 1962), the “stock of research and development” (Romer 1986), “human capital (Lucas 1988), “trade spillovers” (Coe and Helpman 1995; Coe et al. 1997), and “trade openness” (Diao et al. (2005)). However, *if the explanation for the spillover that endogenously determines technology change is the loosening of constraints on the utilization of the technology, then this is just another way of saying that TFP growth is primarily determined by the efficiency with which the existing technology (inclusive of innovations) is utilized* (Sickles et al. 2015).

We will take a reduced form approach in much of what we discuss below. The literature on structural modeling of productivity models is quite dense and, outside the scope of our study. The broader structural modeling of static and dynamic productivity models (see for example, Olley and Pakes 1996) speaks to other issues than those we focus on herein. These issues involve, among other things, the role of errors-in-variables, weak instrument bias, index construction, and stability in panel data modeling of production processes. They have been taken up by a number of researchers. The NBER is particularly well-represented. Studies by Griliches and

Hausman (1986), Stoker et al. (2005), Griliches and Mairesse (1990, 1998), and Griliches and Pakes (1984), Diewert and Deaton (2002), Diewert (2004a,b) are but a few in this extensive literature.

17.3 Decomposition of Economic Growth-Innovation and Efficiency Change Identified by Regression

Regression based approaches to decompose productivity growth into technical change and catch-up (efficiency change) components can be based on the following generic model. Assume that the multiple output / multiple input technology can be estimated parametrically using the output distance function (Caves et al. 1982; Coelli and Perelman 1996). We consider distance or single output production functions that are linear in parameters, such as the linear in logs Cobb-Douglas, translog, generalized-Leontief and quadratic. These constitute the predominant functional forms used in productivity studies. We begin with a relatively simple representation of the output distance function as an m -output, n -input deterministic distance function $D_o(Y, X)$ given by the Young index, described in Balk (2008):

$$D_o(Y, X) = \frac{\prod_{j=1}^m Y_{it}^{\gamma_j}}{\prod_{k=1}^n X_{it}^{\delta_k}} \leq 1. \quad (17.5)$$

The output-distance function $D_o(Y, X)$ is non-decreasing, homogeneous, and convex in Y and non-increasing and quasi-convex in X . The output distance function is linearly homogeneous in outputs. Take logs, add a disturbance term v_{it} to account for nonsystematic error in observations, functional form, etc. and a technical efficiency term $\eta_i(t)$ to reflect the nonnegative difference between the upper bound of unity for the distance function and the observed value of the distance function for country i at time t . Then we can write the distance function as:

$$-y_{1,it} = \sum_{j=2}^m \gamma_j y_{jit}^* + \sum_{k=1}^n \delta_k x_{kit}^* + \eta_i(t) + u_{it} \quad (17.6)$$

where $y_{ jit,j=2,\dots,m}^* = \ln(Y_{ jit}/Y_{1it})$ and $x_{ kit}^* = \ln(X_{ kit})$.

After redefining a few variables the distance function can be written as

$$y_{it} = x_{it}\beta + \eta_i(t) + v_{it}. \quad (17.7)$$

The Cobb-Douglas specification of the distance function (Klein 1953) has been criticized for its assumption of separability of outputs and inputs and for incorrect curvature as the production possibility frontier is convex instead of concave. However, as pointed out by Coelli (2000), the Cobb-Douglas remains a reasonable and parsimonious first-order local approximation to the true function.

The translog output distance function introduces second-order terms that allow for greater flexibility without sacrificing the possibility of proper local curvature and lifts the assumption that outputs and inputs are separable. The translog output distance function also can be framed in this canonical model representation of a linear panel model with country-specific and time-varying heterogeneity. If the translog technology is applied, the distance function takes the form:

$$\begin{aligned} -y_{1it} = & \sum_{j=2}^m \gamma_j y_{j1t}^* + \frac{1}{2} \sum_{j=2}^m \sum_{l=1}^m \gamma_{jl} y_{j1t}^* y_{lit}^* + \sum_{k=1}^n \delta_k x_{kit}^* \\ & + \frac{1}{2} \sum_{k=1}^n \sum_{p=1}^n \delta_{kp} x_{kit}^* x_{pit}^* + \sum_{j=2}^m \sum_{k=1}^n \theta_{jk} y_{j1t}^* x_{lit}^* + \eta_{it} + u_{it} \end{aligned} \quad (17.8)$$

Since the model is linear in parameters then after redefining a few variables the translog distance function also can be written as $y_{it} = x_{it}\beta + \eta_i(t) + v_{it}$.

A similar transparent reparametrization of any distance function that is linear in parameters can be used to estimate other linear in parameters distance or production functions such as the generalized Leontief or quadratic. If the technology involves multiple outputs, then the right hand side endogenous variables must be instrumented. Whether or not the effects need to be instrumented depends on their orthogonality with all or a subset of the regressors. This is the generic model for estimating efficiency change using panel data (frontier) methods that we will explore below. If we assume that innovations are available to all firms and that country- or firm-specific idiosyncratic errors are due to relative inefficiencies, then we can decompose sources of *TFP* growth in a variety of ways. The overall level of innovation change (innovation is assumed to be equally appropriable by all countries) can be measured directly by such factors as a distributed lag of R&D expenditures, or patent activity, or some such direct measure of innovation. The overall level of innovation change also can be proxied by the time index approach of Baltagi and Griffin (1988), linear time trends, or some other type of time variable. Innovation measured in any of these ways would be identified in most empirical settings. Direct measures are identified of course by the assumption that the matrix of regressors has full column rank, and the indirect measures by functional form assumptions. For example, the index number approach used in Baltagi and Griffin is identified by its nonlinear construction. Innovation also is often proxied by exogenous or stochastic linear time trends (Bai et al. 2009).

17.3.1 What is the Correct Model?

One can explore a number of regression-based methods introduced into the literature to measure productivity growth and its decomposition into innovation and catch-up, or efficiency change (Sickles et al. 2015). We will examine estimates from

these various specifications using the generic linear panel data model with time-varying and cross-sectionally varying effects that is given above. The generic panel data model $y_{it} = x_{it}\beta + \eta_i(t) + v_{it}$ nests all multi-output/multi-input panel models that are linear in parameters and can be used to estimate productivity growth and decompose it into innovation and catch-up. We assume that we have a balanced panel although this is done more for notational convenience than for substantive reasons. The generic model of course nests all models that we introduce below for which there is no temporal change in technical efficiency, that is, the usual fixed or random effects stochastic panel frontier models introduced by Pitt and Lee (1981) and Schmidt and Sickles (1984).

We first discuss the most common estimators in use and those that have been introduced rather recently and how they can be implemented in empirical applications. We then show how these methods can be used in a model averaging exercise to evaluate world productivity trends.

17.3.1.1 The Cornwell et al. (1990) Panel Stochastic Frontier Model

Extensions of the panel data model by Cornwell, Schmidt, and Sickles (CSS) 1990 generalized the model in which heterogeneity was only allowed in the intercept by allowing for heterogeneity in slopes as well and this permitted researchers to estimate productivity change that was specific to the cross-sectional unit (firms, industries, countries) that could change over time.

A particular parameterization of the CSS model that accomplishes this objective is based on the assumption that in the generic model above ($y_{it} = x_{it}\beta + \eta_i(t) + v_{it}$) the heterogeneity term $\eta_i(t)$ is given by

$$\eta_i(t) = W_{it}\delta_i + v_{it}.$$

The coefficients in the vector δ_i depend on the different cross-sectional units i and represent heterogeneity in slopes. In their application to the US commercial airline industry CSS specified $W_{it} = (1, t, t^2)$ although this was just a parsimonious parameterization useful for their application. The CSS estimator does not in general limit the effects to be quadratic in time but does restrict the effects to be linear in the parameters of the variables whose slopes vary by cross-sectional units. Three different estimators were derived based on differing assumptions made in regard to the correlation of the efficiency effects and the regressors, specifically relating to the correlation between the error term u and regressors X and W . These estimators are the *within* (FE) estimator, which allows for correlation between all of the regressors and the effects, the *gls* estimator, which is consistent when no correlation exists between the technical efficiency term and the regressors (Pitt and Lee 1981; Kumbhakar 1990), and the *efficient instrumental variables* estimator, which can be obtained by assuming orthogonality of some of the regressors with the technical efficiency effects. The explicit formulas for deriving each estimator and methods

for estimating the δ_i parameters are provided in the Cornwell et al. (1990). Relative efficiencies, normalized by the consistent estimate of the order statistics identifying the most efficient cross-sectional unit, are calculated as:

$$\hat{\eta}(t) = \max_j[\hat{\eta}_j(t)]$$

and

$$RE_i(t) = \hat{\eta}(t) - \hat{\eta}_i(t).$$

Here $RE_i(t)$ is the relative efficiency of the i th cross-sectional unit at time t . For this class of models the regressors X contain a time trend interpreted as the overall level of innovation. When it is combined with the efficiency term $\hat{\eta}_j(t)$ we have a decomposition of TFP into innovation and catch-up. When the time trend and the efficiency term both enter the model linearly then the decomposition is not identified using the within estimator. The composition is identified for the gls and for selected variants of the efficient IV model, such as those used in the Cornwell et al. (1990) airline study. In our empirical illustration using the UNIDO data to estimate world productivity growth that follows, we utilize the gls version of the CSS estimator (labelled CSSG) and the efficiency IV estimator (labelled EIV).

17.3.1.2 The Kumbhakar (1990) Panel Stochastic Frontier Model

Consider the linear in log production function:

$$y_{it} = x_{it}\beta + \eta_i(t) + v_{it}$$

$\eta_i(t) = \gamma(t)\tau_i$, where v_{it} is assumed i.i.d. with distribution $N(0, \sigma_v^2)$. $\eta_i(t)$ is the inefficiency term with a time-varying factor $\gamma(t)$ and time-invariant characteristics τ_i . τ_i is assumed to be distributed as *i.i.d.* half-normal and $\gamma(t)$ is specified as the logistic function

$$\gamma(t) = (1 + \exp(bt + ct^2))^{-1}.$$

Here $\gamma(t)$ is bounded between (0, 1) and accommodates increasing, decreasing or time-invariant inefficiency behavior as the parameters b and c vary. Although the Kumbhakar model also estimates allocative efficiency from side conditions implied by cost-minimization (Schmidt and Lovell 1979) we will only examine the portion of his model that directly pertains to the technical inefficiency/innovation decomposition of productivity change. Parametric maximum likelihood is used for estimation of the main parameters of the model. The inefficiency term is estimated by analogue methods based on the population first moment of $\tau_i|\theta_i$. The best predictor of technical efficiency is then given by $E(\exp\{\gamma(t)\tau_i|\theta_i\})$ and efficiency for each unit is given by $\hat{\eta}_i(t) = \gamma(t)\hat{\tau}_i$.

17.3.1.3 The Battese and Coelli Model (1992, 1995)

The production function is given by the generic model

$$y_{it} = x_{it}\beta + \eta_i(t) + v_{it}. \quad (17.9)$$

The effects are specified as

$$\eta_i(t) = -\{\exp[-\eta(t-T)]\}u_i,$$

where v_{it} is assumed to be an *i.i.d.* $N(0, \sigma_v^2)$ random variable, u_{it} is assumed to follow an *i.i.d.* non-negative truncated $N(\mu, \sigma^2)$ distribution, η is a scalar and the temporal movement of the technical efficiency effects depends on the sign of η . Time invariant technical efficiency corresponds to $\eta = 0$. A richer temporal path for firm efficiency effects can be obtained by specifying $\eta(t-T)$ as

$$\eta_t(t-T) = 1 + a(t-T) + b(t-T)^2.$$

This permits the temporal pattern of technical efficiency effects to be convex or concave rather than simply increasing or decreasing at a constant rate. The model is estimated by parametric mle and the minimum-mean-squared-error predictor of the efficiency for unit i at time t is

$$E[\exp(-u_{it})|\epsilon_i] = \left\{ \frac{1 - \Phi[\eta_{it}\sigma_i^* - (\mu_i^*)/\sigma_i^*]}{1 - \Phi(-\mu_i^*/\sigma_i^*)} \right\} \exp \left[-\eta_{it}\mu_i^* + \frac{1}{2}\eta_{it}^2\sigma_i^{*2} \right].$$

Estimates of technical change due to innovation are based on the coefficient of a time trend in the regression. The effect of innovation as distinct from catch-up is identified by the nonlinear time effects in the linear technical efficiency term and thus the decomposition of *TFP* growth into a technological change and efficiency change component is quite natural with this estimator. Cuesta (2000) generalized (Battese and Coelli 1992) by allowing each country (firm, etc.) to have its own time path of technical inefficiency. Extensions of the Battese and Coelli model that allow for technical inefficiency to be determined by a set of environmental factors that differ from those that determine the frontier itself are given in Battese and Coelli (1995). These were also addressed by Reifschneider and Stevenson (1991) and by Good et al. (1995). Environmental factors that were allowed to partially determine the level of inefficiency and productivity were introduced in Cornwell et al. (1990) and in Good et al. (1993).

17.3.1.4 Alternatives to the Classical Parametric Stochastic Panel Frontier Approaches

Many other variations in the basic panel model treatment of inefficiency have been considered in the literature. We do not pursue those in this here but direct the reader to the work of Park, Sickles and Simar (PSS; 1998, 2003, 2006), who

considered linear stochastic frontier panel models in which the distribution of country specific technical efficiency effects is estimated nonparametrically. The latent class models of Orea and Kumbhakar (2004), Tsionas and Kumbhakar (2004), and Greene (2005b) relate to work on production heterogeneity by Mundlak (1961, 1978) and Griliches (1979), among others. Kneip et al. (2012) assume a linear semiparametric panel frontier that allows for an arbitrary pattern of technical change $\eta_i(t)$ based on a general factor model set-up. Their specification of the effects is more flexible than parametric methods and the multiplicative effects models of Lee and Schmidt (1993), Ahn et al. (2007), Bai (2009), and Bai and Ng (2011). Ahn et al. (2013) generalize Ahn et al. (2007) and consider a panel data model with multiple individual effects that also change over time and focus on large N and finite T asymptotics. Additional estimators that have been proposed for panel stochastic frontiers and that are also quite appropriate for general panel data problems are the *Bayesian Stochastic Frontier Model* (Liu et al. 2013), which builds on earlier work by Van den Broeck et al. (1994) and Tsionas (2006), the *Bounded Inefficiency Model* of Almanidis et al. (2014) and related models of Lee (1996), Lee and Lee (2014), and Orea and Steinbuks (2012) as well as the “*True*” *Fixed Effects Model* of Greene (2005a,b). Kumbhakar et al. (2013) considered a semiparametric smooth coefficient model to estimate the TFP growth of certain production technologies that addresses the *Skewness Problem* in classical SFA modeling considered by Feng, Horrace and Wu (2013), Almanidis and Sickles (2012) and Almanidis et al. (2014). The *Spatial Stochastic Frontier* shows great promise and has been pursued in recent work by Glass et al. (2014, 2015) based on the original contribution by Druska and Horrace (2004). Work on productivity measurement in the presence of spatial heterogeneity has also recently been pursued Mastromarco and Shin (2013), Entur and Musolesi (2015), and Demetrescu and Homm (2013). These are alternatives to less structured approaches to address cross-sectional dependence in panel data models using methods such as those developed by Pesaran (2007). *Factor Models* continue to be pursued in the context of productivity modeling in panel data contexts and the space for such approaches is getting quite dense as pointed out by Kneip and Sickles (2012).

17.4 Can We Combine Model Estimates Instead of Choosing the Best?

Discovering the true model might not be possible. Statistical inference based on the “post-model-selection estimators” (Leeb and Pötscher 2005) might lead to invalid analysis. Different selecting criteria might give contradicted ranking orders and focusing on one model and dismissing the results of alternative specifications may compromise the information content of the information set. As discussed in Burnham and Anderson (2002), if observed data are conceptualized as random variables, the sample variability introduces uncertain inference from the particular

data set. Model selection is a special case of weighting models in which one model is given the entire weight. Combining model estimates and forecasts can be motivated on the basis of economic theory based on models of majority voting and the Tullock contest function. It can also be motivated on the basis of statistical theory via model averaging and forecast combination theory.

Most model averaging methods take on a Bayesian perspective, although many recent studies have a frequentist interpretation. In the frequentist literature, the weights are usually based on AIC or BIC criterion. What we will employ here is the method proposed in Hansen (2007), in which the weights are chosen by minimizing a Mallows criterion. It was shown that the resulting estimator can asymptotically achieve the lowest squared error among a finite number of model averaging estimators. For this application our unrestricted model is

$$y_{it} = \sum_{j=1}^p \beta_j x_{itj} + \sum_{r=0}^{\infty} \delta_{ir} t^r + \epsilon_{it}, \quad i = 1, \dots, n; t = 1, \dots, T.$$

Here ϵ_{it} is the standard noise component assumed to be *iid* $N(0, \sigma)$. Let $z_{it} = (x_{it1}, \dots, x_{itp}, 1, t, t^2, \dots)$ be the vector of all regressors, and $\gamma = (\beta_1, \dots, \beta_p, \delta_0, \delta_1, \dots)$ be the associated parameter vector. Then the model can be compactly expressed as

$$y_{it} = \sum_{j=1}^{\infty} z_{itj} \gamma_j + \epsilon_{it}, \quad (17.10)$$

where z_{itj} is a vector with countably infinite entries and can include regressors that are terms of a series expansion that is linear in parameters.

Denote $\mu_{it} = \sum_{j=1}^{\infty} z_{itj} \gamma_j$, it is assumed $E[\mu_{it}^2] < \infty$ and μ_{it} converges to the mean square. Consider a set of M nested models, and assume the m th model uses the first k_m variables, with $p < k_1 < k_2 < \dots < k_M$. Put in matrix form, the m th model is

$$Y = Z_m \Gamma_m + \epsilon. \quad (17.11)$$

Let $\hat{\Gamma}_m$ be the estimate of the coefficients in the m th model, and let $w = (w_1, \dots, w_M)$ be the associated weight vector for each model, where $w_m \in [0, 1]$ and $\sum_{m=1}^M w_m = 1$. Then the model average estimator of the coefficients for the unrestricted model is

$$\hat{\Gamma} = \sum_{m=1}^M w_m \begin{pmatrix} \hat{\Gamma}_m \\ 0 \end{pmatrix}. \quad (17.12)$$

We further define $k(w) \equiv \sum_{m=1}^M w_m k_m$, then the Mallows criterion can be stated as

$$C(w) = (Y - Z_M \hat{F})'(Y - Z_M \hat{F}) + 2\sigma^2 k(w) \quad (17.13)$$

and an optimal weight is obtained by numerically minimizing $C(w)$.

The unrestricted model we consider is from Kneip et al. (2012) and is specified as

$$y_{it} = \beta_0(t) + \sum_{j=1}^p \beta_j x_{ij} + u_{it} + \epsilon_{it}, \quad i = 1, \dots, n; t = 1, \dots, T. \quad (17.14)$$

The u_{it} 's are assumed to be smooth time-varying individual effects where $\sum_i^n u_{it} = 0$, $t = 1, \dots, T$. The individual effects are assumed to be affected by a set of underlying factors and are model by linear combinations of some basis functions.

$$u_{it} = \sum_{r=1}^L \delta_{ir} v_r(t) \quad i = 1, \dots, n \quad (17.15)$$

where $\beta_0(t)$ is some average function and can be eliminated by transforming the model to the centered form:

$$y_{it} - \bar{y}_t = \sum_{j=1}^p \beta_j (x_{ij} - \bar{x}_{ij}) + \sum_{r=1}^L \delta_{ir} v_r(t) + \epsilon_{it} - \bar{\epsilon}_i, \quad i = 1, \dots, n; t = 1, \dots, T, \quad (17.16)$$

where $\bar{y}_t = \frac{1}{n} \sum_i y_{it}$, $\bar{x}_{ij} = \frac{1}{n} \sum_i x_{ij}$ and $\bar{\epsilon}_i = \frac{1}{n} \sum_i \epsilon_{it}$. Denote $\tilde{y}_{it} = y_{it} - \bar{y}_t$ and $\tilde{x}_{ij} = x_{ij} - \bar{x}_{ij}$.

The functional form used for estimation can be written as

$$\tilde{y}_{it} = \sum_{j=1}^p \beta_j \tilde{x}_{ij} + \sum_{r=1}^L \delta_{ir} v_r(t) + \tilde{\epsilon}_{it}, \quad i = 1, \dots, n; t = 1, \dots, T \quad (17.17)$$

This model nests several specifications in stochastic frontier analysis. When $v_r(t) = t^{r-1}$ and $L = 3$ we have the Cornwell et al. (1990) model discussed above. To show how Kumbhakar (1990) is nested in the general model consider a translog production function that is linear in parameters and can be expressed as

$$y_{it} = X'_{it} \beta + u_{it} + \epsilon_{it}, \quad (17.18)$$

where the u_{it} 's represent the individual effects and given by $u_{it} = v(t)\theta_i = (1 + \exp(bt + ct^2))^{-1}\theta_i$. Taking a Taylor expansion of $v(t)$ at $t = 0$, the individual effects can be expressed as

$$\begin{aligned} u_{it} &= \sum_{r=0}^{\infty} \frac{v^{(r)}(0)}{r!} t^r \theta_i \\ &= \theta_i \left(\frac{1}{2} - \frac{1}{4} bt + \frac{1}{4} ct^2 + \dots \right). \end{aligned} \quad (17.19)$$

With a finite time period under study, the exponential time-varying path can be closely approximated by a polynomial function of finite degree L_1 . Thus the model can be written as

$$y_{it} = X'_{it} \beta + \sum_{r=0}^{L_1} \delta_{ir} t^r + \epsilon_{it} \quad (17.20)$$

with $\delta_{ir} = \theta_i \frac{v^{(r)}(0)}{r!}$.

The Battese and Coelli (1992) model can also be nested in the KSS general model using a Taylor expansion. The basic setting is the same, while the individual effects are assumed to follow a different time-varying path.

$$u_{it} = -\eta_{it} u_i = -\{\exp[-\eta(t-T)]\} u_i.$$

Taking a Taylor expansion of this function, we have

$$\begin{aligned} u_{it} &= -\sum_{r=0}^{\infty} \frac{\eta^{(r)}(0)}{r!} t^r u_i \\ &= -e^{\eta T} u_i + \eta e^{\eta T} u_{it} - \frac{1}{2} \eta^2 e^{\eta T} u_i + \dots \end{aligned} \quad (17.21)$$

This exponential function can be sufficiently well approximated by a polynomial function of finite degree L_2 in empirical studies. Setting $\delta_{ir} = -u_i \frac{\eta^{(r)}(0)}{r!}$, the model can be written as

$$y_{it} = X'_{it} \beta + \sum_{r=0}^{L_2} \delta_{ir} t^r + \epsilon_{it} \quad (17.22)$$

The KSS model also nests the traditional **random** and **fixed effects** estimators as these are special cases of the CSS estimator.

In the next section we utilize our model averaging methods on these various nested special cases of the general KSS specification and analyze average productivity growth rates and their decomposition into efficiency change and innovation change across various countries in the world economy. The section illustrates the feasibility of the approach and the potential gains that researchers can derive from bringing different models and different assumptions on which they are based to bear in analyzing an important determinant of economic growth and long term economic welfare of a country and of the world economy.

17.5 Taking Model Averaging to the Data-Some Preliminary Results in a Study of World Productivity (1960–2010)

17.5.1 UNIDO Data Description

The World Productivity Database (WPD) provides information on measures of the level and growth of TFP based on 12 different empirical methods across 112 countries over the period 1960–2010. The principal data source is the Penn World Tables from which (chain weighted) GDP and investment are obtained, both in purchasing power parity (1996) US dollars. From the Groningen Growth and Development Centre and Asian Development Bank (ADB, various issues), data on employment and hours worked were also obtained. Unemployment rates and key indicators of the labor market were collected from the International Labor Organization (ILO) Yearbook, and ADB (various issues). Various capital input measures were also constructed. Capital (K) is arguably the most difficult production factor to measure. The WPD presents 4 different approaches based on: (1) different computations for the initial capital stock (2) the depreciation rate (3) schedule for depreciation, and (4) the lifetime of the asset. The different capital measures are labeled K06, K13, K_s and K_{eff} . Common to the first three capital measures is that capital is assumed to depreciate at a constant rate over time. The first two capital stocks differ only in terms of their assumed depreciation rates (6 % and 13.3 %, respectively, which correspond to about 12 and 6 year asset lives). The different depreciation rates emphasize the importance of either the initial capital or the effect of recent investments. K06 and K13 are based on assumption that ten years of investment serve as an adequate proxy for the initial capital stock K0. Another common way of computing the initial capital stock is to assume that the country is at its steady state capital-output ratio. This leads to a level of steady-state capital service flows (K_s) from a capital stock whose assumed depreciation rate is 6 % per year. A different way of measuring capital focuses on the profile of capital productivity and utilizes a time-varying depreciation rate. As the asset ages, its capital services decline at an increasing rate. This leads to the measure labelled K_{eff} . There also are two kinds of labor utilization rates for which labor force can be adjusted in our analysis. One is based on variations in the numbers employed and one is based on variations in hours worked. The first alternative to labor force (LF) is employment (EMP), which is obtained as a direct measure of employment. The second is derived by applying unemployment rates to LF data which leads to derived employment (DEMP).

We apply our model averaging methodology to the OECD (24 countries) based on UNIDO data from 1960 to 2010. The countries in the **OECD** are: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Greece, Iceland, Ireland, Italy, Japan, Republic of Korea, Luxembourg, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, UK, and USA.

17.5.1.1 Summary of Preliminary TFP Findings for the OECD

We choose K06, K13 and Keff as the capital inputs. We use EMP as the labor input. The observation periods are from 1960 to 2010. We use the CSS, K, BC, RE, FE specifications using EMP and three different capital measures. Thus, for this exercise we have 15 different sets of estimates. We aggregate the results by country, by time, to construct aggregate summary measures of technical innovation and technical efficiency growth over the 50 years in our sample of OECD countries. Aggregation is based on utilize geometric means using exchange and ppp weighted gdp shares by each country as the weights. Results suggest that the impact of catch-up relative to technical innovation is marginal. Preliminary results based on from our model averaging exercise yields

$$\begin{aligned}
 & \text{OECD} \\
 & 1960-2010 \\
 \text{TFP growth} & = 1.04\% \text{ (innovation)} + .09\% \text{ (catch-up)} \\
 & = 1.13\%.
 \end{aligned}$$

17.6 Conclusion

We have discussed different theories on economic growth and productivity measurement and the econometric specifications they imply. We develop a variety of methodologies to combine the results from different models. Our methodologies are illustrated with date from the World Productivity Database gathered by UNIDO. TFP growth is decomposed to two components: technical efficiency change and technological change. We aggregate growth rates of different efficiency measures using model averaging criteria. We find out that in the time period between 1960 and 2010, OECD countries averaged about 1 % TFP growth. Innovation that expanded the production possibility frontier plays a much more significant role than catch-up in improving TFP.

References

- Almanidis P, Sickles RC (2012) The skewness problem in stochastic frontier models: fact or fiction? In: Van Keilegom I, Wilson P (ed) Exploring research frontiers in contemporary statistics and econometrics: a festschrift in honor of Leopold Simar. Springer, New York
- Alminidis P, Qian J, Sickles RC (2014) Stochastic frontiers with bounded inefficiency. Sickles RC, Horrace WC (ed) Chapter 3 of Festschrift in honor of Peter Schmidt: econometric methods and applications. Springer Science & Business Media, New York, NY, 47–82.
- Ahn SC, Lee YH, Schmidt P (2007) Stochastic frontier models with multiple time-varying individual Effects. *J Prod Anal* 27(1):1–12

- Ahn SC, Lee YH, Schmidt P (2013) Panel data models with multiple time-varying individual effects. *J Econ* 174(1):1–14
- Arrow KJ (1962) The economic implications of learning by doing. *Rev Econ Stud* 29(3):155–173
- Bai J (2009) Panel data models with interactive fixed effects. *Econometrica* 77(4):1229–1279
- Bai J, Ng S (2011) Principal components estimation and identification of the factors. Department of Economics, Columbia University
- Bai J, Kao C, Ng S (2009) Panel cointegration with global stochastic trends. *J Econ* (149):82–99
- Balk BM (2008) Price and quantity index numbers models for measuring aggregate change and difference. Cambridge University Press, Cambridge
- Baltagi BH, Griffin JM (1988) A general index of technical change. *J Polit Econ* 96:20–41
- Battese GE, Coelli TJ (1992) Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India. *J Prod Anal* 3(1–2):153–169
- Battese GE, Coelli TJ (1995) A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empir Econ* 20:325–332
- Burnham KP, Anderson DR (2002) Model selection and multi-model inference: a practical information-theoretic approach. Springer, New York
- Caves DW, Christensen LR, Diewert WE (1982) The economic theory of index numbers and the measurement of input, output, and productivity. *Econometrica: J Econ Soc* 50:1393–1414
- Coe DT, Helpman E (1995) International R&D spillovers. *Eur Econ Rev* 39(5):859–887
- Coe DT, Helpman E, Hoffmaister A (1997) North-south R&D spillovers. *Econ J* 107(440): 134–149
- Coelli T (2000) Econometric estimation Of the distance function representation of a production technology. CEPA, University of New England
- Coelli T, Perelman S (1996) Efficiency measurement, multiple output technologies and distance functions: with application to European railways. *Eur J Oper Res* 117:326–339
- Cornwell C, Schmidt P, Sickles RC (1990) Production frontiers with cross-sectional and time-series variation in efficiency levels. *J Econ* 46(1):185–200
- Cuesta RA (2000) A production model with firm-specific temporal variation in technical inefficiency: with application to Spanish dairy farms. *J Prod Anal* 13(2):139–158
- Demetrescu M, Homm U (2013) A directed test of no cross-sectional error correlation in large-N panel data models. Working paper. University of Bonn, Germany
- Diewert E (2004a) Preface, (with Hill P, Armknecht P); Chapter 15, basic index number theory; Chapter 16, the axiomatic and stochastic approaches to index number theory; Chapter 17, the economic approach to index number theory: the single household case; Chapter 18, the economic approach to index number theory: the many household case; Chapter 19, price indices using an artificial data set; Chapter 20, elementary indices; Chapter 22, the treatment of seasonal products. In: Consumer price index manual: theory and practice. International Labour Organization, Geneva
- Diewert E (2004b) Preface, (with Hill P, Armknecht P); Chapter 15, basic index number theory; Chapter 16, the axiomatic and stochastic approaches to index number theory; Chapter 17, economic approach; Chapter 18, aggregation issues; Chapter 19, price indices using an artificial data set; Chapter 20, elementary indices; Chapter 21, quality change and hedonics; Chapter 22, treatment of seasonal products. In: Producer price index manual: theory and practice. International Monetary Fund, Washington DC
- Diewert and Deaton E, Deaton A (2002) Conceptual foundations of price and cost of living indexes. In: Schultze CL, Mackie C (ed) At what price? Conceptualizing and measuring cost of living and price indexes. National Academy Press, Washington
- Diao X, Rattsø J, Stokke HE (2005) International spillovers, productivity growth and openness in Thailand: an intertemporal general equilibrium analysis. *J Dev Econ* 76(2):429–450
- Druska V, Horrace WC (2004) Generalized moments estimation for spatial panel data: Indonesian rice farming. *Am J Agric Econ* 86: 185–198
- Ertur C, Musolesi A (2015) Weak and strong cross-sectional dependence: a panel data analysis of international technology diffusion. No. 0415. SEEDS, Sustainability Environmental Economics and Dynamics Studies

- Feng Q, Horrace W, Wu GL (2013) Wrong skewness and finite sample correction in parametric stochastic frontier models. Mimeo
- Feng Q, Horrace WC, Wu GL (2013) Wrong skewness and finite sample correction in parametric stochastic frontier models. Center for Policy Research Working Paper 154. Syracuse University.
- Fisher I (1927) The making of index numbers: a study of their varieties, tests, and reliability. Houghton Mifflin company, Boston
- Glass A, Kenjegalieva L, Sickles RC (2015) A spatial autoregressive stochastic frontier model for panel data with asymmetric efficiency spillover. *Journal of Econometrics* (2015). <http://dx.doi.org/10.1016/j.jeconom.2015.06.011>
- Glass A, Kenjegalieva L, Sickles RC (2014) Estimating efficiency spillovers with state level evidence for manufacturing in the US. *Economics Letters* 123.2: 154–159.
- Good DH, Nadiri MI, Roeller LH, Sickles RC (1993) Efficiency and productivity growth comparisons of European and U.S. air carriers: a first look at the data. (In: Mairesse J, Griliches Z (ed) *Journal of Productivity Analysis Special Issue*) *J Prod Anal* 4:115–125
- Good DH, Roeller LH, Sickles RC (1995) Airline efficiency differences between Europe and the U.S.: implications for the pace of E.C. integration and domestic regulation. *Eur J Oper Res* 80:508–518
- Good DH, Nadiri MI, Sickles RC (1997) Index number and factor demand approaches to the estimation of productivity. In: Pesaran MH, Schmidt P (ed) *Handbook of Applied Economics*, Volume II-Microeometrics. Oxford, Basil Blackwell
- Greene W (2005a) fixed and random effects in stochastic frontier models. *J Prod Anal* 23(1):7–32
- Greene W (2005b) reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *J Econ* 126(2):269–303
- Griliches Z (1957) Hybrid corn: an exploration in the economics of technological change. *Econometrica*: *J Econ Soc* 25(4):501–522
- Griliches Z (1979) Issues in assessing the contribution of research and development to productivity growth. *Bell J Econ* 92–116
- Griliches Z, Hausman JA (1986) Errors in variables in panel data. *J Econ* 31:93–118
- Griliches Z, Pakes A (1984) Distributed lags in short panels with an application to the specification of depreciation patterns and capital stock constructs. *Rev Econ Stud* 51(2): 1175–1189
- Griliches Z, Mairesse J (1990) Heterogeneity in panel data: are there stable production functions? *Essays in Honor of Edmond Malinvaud*. vol 3, MIT Press, Cambridge, pp. 192–231
- Griliches Z, Mairesse J (1998) Production functions: the search for identification. In: Ström S (ed) *Econometrics and Economic Theory in the 20th Century: The Ragnar Frish Centennial Symposium*. Cambridge University Press, Cambridge
- Hansen BE (2007) Least squares model averaging. *Econometrica* 75(4):1175–1189
- Hultberg PT, Nadiri MI, Sickles RC (1999) An international comparison of technology adoption and efficiency: a dynamic panel model. *Annales d'Economie et de Statistique* 449–474
- Hultberg PT, Nadiri MI, Sickles RC (2004) Cross-country catch-up in the manufacturing sector: impacts of heterogeneity on convergence and technology adoption. *Empir Econ* 29(4):753–768
- Jorgenson D, Griliches Z (1972) Issues in growth accounting: a reply to Edward F. Denison. *Survey of Current Business* 52.5, Part II
- Kendrick JW (1961) Front matter, productivity trends in the United States. *Productivity Trends in the United States* 52–0. NBER
- Kim JI, Lau LJ (1994) The sources of economic growth of the east Asian newly industrialized countries. *J Jpn Int Econ* 8(3):235–271
- Klein L (1953) A textbook of econometrics. Evanston [etc.], Peterson, New York
- Kneip A, Sickles RC (2012) Panel data, factor models, and the solow residual. In: Van Keilegom I, Wilson P (ed) *Exploring research frontiers in contemporary statistics and econometrics: a festschrift in honor of Leopold Simar*. Springer, New York
- Kneip A, Sickles RC, Song W (2012) A new panel data treatment for heterogeneity in time trends. *Econ Theory* 28: 590–628
- Kumbhakar SC, Parmeter C, Tsionas EG (2013) A zero inefficiency stochastic frontier model. *J Econ* 172(1):66–76
- Krugman P (1994) The myth of Asia's miracle. *Foreign Affairs* 73:62–78

- Kumbhakar SC (1990) Production frontiers, panel data, and time-varying technical inefficiency. *J Econ* 46(1):201–211
- Lee Y (1996) Tail truncated stochastic frontier models. *J Econ Theory Econ* 2:137–152
- Lee Y, Lee S. (2014) Stochastic frontier models with threshold efficiency. *Journal of Productivity Analysis* 42.1: 45–54.
- Leeb H, Pötscher B M (2005) Model selection and inference: facts and fiction. *Econ Theory* 21(1):21–59
- Lee YH, Schmidt P (1993) A production frontier model with flexible temporal variation in technical efficiency. *The measurement of productive efficiency: Techniques and applications*, Oxford University Press, New York, pp 237–255
- Liu J, Sickles RC, Tsionas E (2013) Bayesian treatment to panel data models with time-varying heterogeneity. Working paper. Rice University.
- Lucas RE Jr (1988) On the mechanics of economic development. *J Monet Econ* 22(1):3–42
- Mansfield E (1961) Technical change and the rate of imitation. *Econometrica* 29(4):741–766
- Mastromarco C, Shin Y (2013) Modelling technical efficiency in cross sectionally dependent panels. Working paper. University of Lecce, Italy
- Mundlak Y (1961) Empirical production function free of management bias. *J Farm Econ* 43(1): 44–56
- Mundlak Y (1978) On the pooling of time series and cross section data. *Econometrica: J Econ Soc* 69–85
- Olley GS, Pakes A (1996) The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64(6):1263–1297
- Orea L, Kumbhakar SC (2004) Efficiency measurement using a latent class stochastic frontier model. *Empir Econ* 29(1):169–183
- Orea L, Steinbuk J (2012) Estimating market power in homogenous product markets using a composed error model: application to the California electricity market. University of Cambridge, Faculty of Economics
- Park BU, Sickles RC, Simar L (1998) Stochastic panel frontiers: a semiparametric approach. *J Econ* 84(2):273–301
- Park BU, Sickles RC, Simar L (2003) Semiparametric-efficient estimation of AR (1) panel data models. *J Econ* 117(2):279–309
- Park BU, Sickles RC, Simar L (2006) Semiparametric efficient estimation of dynamic panel data models. *J Econ* 136(1):281–301
- Pesaran MH (2007) A simple panel unit root test in the presence of cross-section dependence. *J Appl Econ* 22:265–312
- Pitt MM, Lee LF (1981) The Measurement and sources of technical inefficiency in the Indonesian weaving industry. *J Dev Econ* 9(1):43–64
- Reifschneider D, Stevenson R (1991) Systematic departures from the frontier: a framework for the analysis of firm inefficiency. *Int Econ Rev* 32:715–723
- Romer PM (1986) Increasing returns and long-run growth. *J Polit Econ* 1002–1037
- Scherer FM (1971) Industrial market structure and economic performance. Rand McNally, Chicago
- Schmidt P, Lovell CK (1979) Estimating technical and allocative inefficiency relative to stochastic production and cost frontiers. *J Econ* 9(3):343–366
- Schmidt P, Sickles RC (1984) Production frontiers and panel data. *J Bus Econ Stat* 2(4):367–374
- Sickles RC, Hao J, Shang C (2014) Panel data and productivity measurement: an analysis of Asian productivity trends. *J Chin Econ Bus Stud* 12(3):211–231
- Sickles RC, Hao J, Shang C (2015) Productivity and panel data. *Oxford handbook of panel data*. Oxford University Press, New York
- Shang C (2015) Reduced form analysis of world productivity growth: a model averaging approach. Mimeo.
- Solow RM (1957) Technical change and the aggregate production function. *Rev Econ Stat* 39(3):312–320
- Smolny W (2000) Sources of productivity growth: an empirical analysis with German sectoral data. *Appl Econ* 32(3):305–314

- Stoker TM, Berndt ER, Ellerman AD, Schennach SM (2005) Panel data analysis of U.S. coal productivity. *J Econ* 127:131–164
- Tsionas EG (2006) Inference in dynamic stochastic frontier models. *J Appl Econ* 21(5):669–676
- Tsionas EG, Kumbhakar SC (2004) Markov switching stochastic frontier model. *Econ J Roy Econ Soc* 7(2):398–425
- Van den Broeck J, Koop G, Osiewalski J, Steel MF (1994) Stochastic frontier models: a bayesian perspective. *J Econ* 61(2):273–303
- Young A (1992) A tale of two cities: factor accumulation and technical change in Hong Kong and Singapore. In: NBER macroeconomics annual 1992, vol 7. MIT press, Massachusetts
- Young A (1995) The tyranny of numbers: confronting the statistical realities of the east Asian growth experience. *Q J Econ* 110(3):641–680

Index

A

- Additive general error model (AGEM), 52–54
Agricultural growth, Sub-Saharan Africa, 175–176
classification, 199
conceptual framework and approach, 178–179
growth accounting approach, 179–180
hybrid approach, 182–183
non-parametric approach, 180–182
decomposition, 178
empirical model and implementation, 183–184
efficiency estimates, 184–185
input elasticities and Cobb-Douglas production function, 186–188
growth and performance, TFP, 188–193
productivity levels and implications, growth ANOVA, 196, 197
appropriate technology, 196
contribution, 195
input and output per worker, 194
technologies, 177
TFP, 176–177
Allen-Uzawa elasticities, 262
Arithmetic weighted averages, 307

B

- Bayesian stochastic frontier model, 314
Bayh-Doyle Act of 1980, 206
Bosman ruling. *See* Player productivity
Bounded inefficiency model, 314

C

- Canadian Productivity Accounts (CPA), 213
Canadian taxfiler data, ICT shortages
average T4 income, 153, 154
cohort approach and analysis, 153–155
immigrant cohort, 155–156
immigrant data, 156–157
longitudinal aspect, 151–153
percentage change, 153
Cobb-Douglas production function, 73–77, 186–188
Common correlated effects (CCE), 187
Concavity violation, 273
Constant returns to scale (CRS/CRTS), 242, 261
Contest theory, 294
Cornwell, Schmidt, and Sickles (CSS) model, 311–312

D

- Data Envelopment Analysis (DEA), 71, 233
CRS, 242
frontier, average data, 242
input, 239–241
output, 241–242
Diewert's contribution
aggregate labour
definition, 2–3
economy wide labour productivity, 3
growth factors, 5
industry n labour productivity, 3
real output price, 3–4

- Diewert's contribution (*cont.*)
 - share of labour, 4
 - value added/output share, 4
 aggregate multifactor
 - economy wide input price index, 7
 - economy wide real input, 7
 - economy wide real output, 6
 - real input price, 7
 - real output price, 6
 TFP, 6–8
- Distribution system operators (DSOs), 233
 - allowed revenue, 236–237
 - revenue caps, 237–238
- Dual-level efficiency approach. *See* National Health Service (NHS) pathology
- Dual level stochastic frontier (DLSF) model, 124
- Dun & Bradstreet (D&B) database, 205
- Dynamic technical efficiency
 - banks, productivity and dynamics, 105–106
 - contributions, 99–100
 - methods, 100–102
 - simulations, 103–105
 - unbiased estimates and confidence set, 102–103
- E**
- Economic performances
 - examination, 273
 - price and Allen partial elasticities
 - capital-energy elasticity, 276
 - consumer goods, 274, 276
 - equipment goods, 274, 275
 - exponential technical progress, 277
 - intermediate goods, 274, 277
 - stochastic specification, 277
 - temporal evolutions of elasticities
 - capital-energy, 279, 280
 - energy-energy, 279, 281
 - labor-energy, 279, 281
 - temporal variability, 280
 - variations of elasticities, 277–279
- Economy wide input price index, 7
- Ecuador
 - data, 111
 - distributional analysis, 115–116
 - emerging economies, 110
 - labour productivity decomposition
 - dynamic strength components and scree plot, 113–115
 - FHK, 111–113
 - labour productivity growth, source of, 110
- Empirical methodology, 164–167
- Entrepreneurship, 203–204
- EUKLEMS data, 163–164
- Exogenous regressors, 105
- F**
- Firm layoffs
 - empirical methodology, 164–167
 - estimation, 171–172
 - EUKLEMS data, 163–164
 - industry shutdown rates, effects of, 163
 - LWF, 162–164
 - probability of, 167–170
 - ways of, 162
- Firms invest, 205
- Four-component DLSF, 126–127
- G**
- General error models
 - AGEM, 53–54
 - fisheries studies and data, 59–60
 - MGEM (*see* Multiplicative general error model (MGEM))
 - operation days, marginal effects, 68–69
 - single equation approach, 60–61
 - standard revenue function, 58–59
 - technical change, 63–64
 - technical efficiency, 65–67
 - Geometric weighted averages, 307
 - Globalization, 294
 - Growth accounting approach, 179–180
- H**
- Hicksian neutral production function, 214–215
- Hotelling's rule, 214, 223
- I**
- Indirect confidence set inference (ICSI)
 - approach, 103
- Indirect inference estimator (IIE), 102
- Information and communications technology (ICT) shortages, 146
- Canadian taxfiler data
 - average T4 income, 153, 154
 - cohort approach and analysis, 153–155
 - immigrant cohort, 155–156
 - immigrant data, 156–157
 - longitudinal aspect, 151–153
 - percentage change, 153

- labour market, indicators of, 148–152
 literature, 146–147
- Innovative startup firms. *See* Knowledge spillovers
- K**
- Kauffman firm survey (KFS), 204–205
- KL-VA production model, 20
- Knowledge spillovers
 data and empirical methodology, 204–206
 intensive margin, 207, 208
- L**
- Labour productivity decomposition
 Diewert's aggregate
 definition, 2–3
 economy wide labour productivity, 3
 growth factors, 5
 industry n labour productivity, 3
 real output price, 3–4
 share of labour, 4
 value added/output share, 4
- Ecuador
 dynamic strength components and scree plot, 113–115
 FHK, 111–113
 geometric means, 11–12
 logarithms of, 9–10
 Taylor series approximation, 9
- Laspeyres-type measure, 30
- Log-quadratic model, flexibility
 estimation, cost function, 284
 parameter estimation, 284, 285
 parameterization, 283, 285
 quality of adjustment, 285–287
 sensitivity, 287
 unit cost, 287, 288
- Longitudinal Administrative Databank (LAD)
 average T4 income, 153, 154
 cohort approach and analysis, 153–155
 immigrant cohort, 155–156
 immigrant data, 156–157
 longitudinal aspect, 151–153
 percentage change, 153
- Longitudinal Employment Analysis Program (LEAP) database, 164
- Longitudinal Worker File (LWF), 162–164
- M**
- Monotonicity property, 41–42, 272–273
- Multifactor productivity growth decomposition
 Diewert's aggregate
 economy wide input price index, 7
 economy wide real input, 7
 economy wide real output, 6
 real input price, 7
 real output price, 6
 TFP, 6–8
 geometric means, 11–12
 logarithms of, 11
 Taylor series approximation, 10–11
- Multiplicative general error model (MGEM)
 FOCs, 55
 inefficiency, 57–58
 log model, linear homogeneity, 55
 normalized quadratic revenue function, 54–55
 revenue function, 55, 56
 revenue maximization model, 55
 revenue model estimation, 61–62
- Mundlak-transformed DLSF, 124–125
- Mundlak-transformed four components DLSF, 127–129
- N**
- National Health Service (NHS) pathology data, 130–131
- DLSF model, 122–124
 efficiency, 128
 four-component DLSF, 126–127
 health policy, implications, 137–139
 modelling multi-level data structures, implications, 139–140
 model selection, 133–136
 multi-level efficiency analysis, 120
- Mundlak-transformed DLSF, 124–125
- Mundlak-transformed four components DLSF, 127–129
 parameter estimates, 131–133
 persistent inefficiency, 121
 SHA, laboratory level and efficiency predictions, 136–137
 unobservable heterogeneity, accounting for, 124
- National Occupation Code (NOC) classification, 149
- Natural capital
 commodity (asset) level data, 225
 energy resources, 212
 Hotelling's rule, 214, 223
 MFP growth, 212, 229
 resource rent and reserve value, 227–228
 resource reserve and extraction, 226–227
 stock and input, 228–229

- Natural capital (*cont.*)
 subsoil mineral (*see* Subsoil resources)
 value-added growth, 230
- Natural gas liquids, 26
- Net Present Value (NPV), 221–223
- Non-parametric approach, 180–182
- North American Industrial Classification System (NAICS), 148–152
- Norwegian regulatory model, 234
 core elements of, 235
 cost norm, 238–239
 calibration of, 244
 DEA model (*see* Data Envelopment Analysis (DEA))
 operational/environmental environments, correction, 243–244
- DSOs, 233
 allowed revenue, 236–237
 revenue caps, 237–238
- Norwegian Water Resources and Energy Directorate (NVE), 233, 243
- O**
- OECD's Regional Patent Database, 205
- Olley-Pakes decomposition, 42–45
- Ordinary least squares (OLS), 72
- P**
- Paasche-type measure, 31
- Panel stochastic frontier model
 CSS model, 311–312
 Kumbhakar model, 312
- Player productivity
 data set and summary
 Bosman ruling, 295
 information gathering, 295
 player entries and source, 296
 statistics, certain variables, 296, 297
- English Premier League, 292–293
- Labour supply effect, 293
- literature review and motivation, 294–295
- post-Bosman period, 292
- RD design
 bin means, Apps, 297, 298
 bin means for MPG, 297, 298
 discontinuous jump, 300, 302
 existence and size estimation, 297
 motivating theory, 297–298
 post- vs. pre-Bosman entrants, 300, 301
 set of assumptions, 299
 treatment effects, 298–300
 treatment parameter, 300
- Production response
 Allen elasticities, substitution, 80
 Cobb-Douglas production function, 76–77
 data, 74–76
 DEA, 71
 efficiency scores, 81, 82
 marginal products and elasticities, 73–74, 79–80
 quantile regression, 72–73, 78–79
 SFA, 71
 stochastic frontier translog production function, 78
- Productivity change
 accounting identities, 18–20
 aggregate labour productivity, 16
 arithmetic means, 18
 bottom-up approach, 17
 complications, 17
 continuing, entering and exiting production units, 20–21
 data accessible, 27–28
 decompositions, 28–29
 geometric and harmonic approach, 40
 intra-unit productivity and Laspeyres-type measures, 31–32
 Laspeyres-type measure, 30–31
 Paasche-type measure, 31
 provisional evaluation, 39–40
 symmetric Bennet-type method, 35–38
 TRAD, CSLS and GEA, 32–35
- enterprise level, 17
- indices, 22–23
- indices and levels, 21
- levels, 24–26
- linking levels and indices, 26–27
- monotonicity, 41–42
- Olley-Pakes decomposition, 42–45
- sectoral shiftshare analysis, 15–16
- top-down approach, 17
- weights, choice of, 45–46
- Productivity growth
 decompositions, 1–2
 Diewert's contribution (*see* Diewert's contribution)
 labour productivity puzzles, 9–10
 multifactor productivity puzzles, 10–12
 index numbers, 12
 measurement (*see* Productivity growth measurement)
- Productivity growth measurement
 model selection
 AIC/BIC criterion, 315
 Battese and Coelli model, 317
 functional form, 316

- Mallows criterion, 315–316
 model averaging methods, 315
 random and fixed effects, 317
 Taylor expansion, 316–317
 uncertain inference, 314–315
 unrestricted model, 315, 316
- regression methods
 Battese and Coelli model, 313
 classical parametric stochastic panel frontier approach, 313–314
 Cobb-Douglas specification, 309
 distance function, 309, 310
 generic panel data model, 311
 multiple output/multiple input technology, 309
 panel stochastic frontier model, 311–312
 parameter estimation, 310
- sources of economic growth
 neoclassical model, 308–309
 productivity efficiency, 306
 TFP, 307
- UNIDO data description
 capital measurement, 318
 labor utilization rates, 318
 TFP findings, OECD, 319
 WPD, 318
- Public capital stock
 empirical application
 data description, 90–91
 findings, 91–95
 spatial weights matrices, specifications, 90
 infrastructures, 83
 investment, economic impact, 84
 QML estimation, 88–89
 spatial autoregressive stochastic frontier model, 84–88
 spillover effects, 84, 85
- Q**
 Quality of fit, 282–284
 Quantile regression, 72–73, 78–79
 Quasi-maximum likelihood (QML) estimation, 85, 88–89
- R**
 Record of Employment (ROE), 164
 Regression discontinuity (RD) design
 bin means, Apps, 297, 298
 bin means for MPG, 297, 298
 discontinuous jump, 300, 302
- existence and size estimation, 297
 motivating theory, 297–298
 post- vs. pre-Bosman entrants, 300, 301
 set of assumptions, 299
 treatment effects, 298–300
 treatment parameter, 300
- Regulatory asset base (RAB), 240
- Resource rent
 and reserve value, 227–228
 subsoil resources, 216–217
 benchmarking, 219–220
 commodity level, 217–218
 decomposition, 220–221
 industry level, 218–219
- Revenue maximization, 59
- Romer model, 308
- S**
 Sectoral shiftshare analysis, 15–16
 Seemingly unrelated regression (SUR) system, 269, 271
 Simple value-added based labour productivity level, 26
 Skewness problem, 314
 Spatial autoregressive stochastic frontier model, 84–88
 Spatial stochastic frontier, 314
 Spillover hypothesis, 294
 Stochastic frontier analysis (SFA), 71
 Strategic Health Authorities (SHAs), 120–121
 Sub-Saharan Africa's (SSAs) agriculture, 175–176
 classification, 199
 conceptual framework and approach, 178–179
 growth accounting approach, 179–180
 hybrid approach, 182–183
 non-parametric approach, 180–182
 decomposition, 178
 empirical model and implementation, 183–184
 efficiency estimates, 184–185
 input elasticities and Cobb-Douglas production function, 186–188
 growth and performance, TFP, 188–193
 productivity levels and implications, growth ANOVA, 196, 197
 appropriate technology, 196
 contribution, 195
 input and output per worker, 194
 technologies, 177
 TFP, 176–177

- Subsoil resources
 estimation, 225–226
 Hicksian neutral production function, 214–215
 industry level measures, 224–225
 logarithmically differentiating yields, 215
 measuring resource rent, 216–217
 benchmarking, 219–220
 commodity level, 217–218
 decomposition, 220–221
 industry level, 218–219
- MFP growth, 229
 parameters, 213
 reserves, 221–223
 resource rent and reserve value, 227, 228
 resource reserve and extraction, 226–227
 stock and input, 228–229
 value-added growth, 230
- Survey of Employment, Payroll and Hours (SEPH), 148, 150
- Symmetric Bennet-type method, 35–38
- T**
- T1 files, 164
- Total factor productivity (TFP), 6–8, 176–177, 307
 findings, OECD, 319
 growth and performance, 188–193
- Translog cost share model
 estimation and empirical results
 economic performances (*see* Economic performances)
 French manufacturing (1970–1989), 269, 270
 hypothesis of symmetry, 271
 quality of fit, 282–284
 regularity conditions, 272–273
 SUR system, 269, 271
 time series data, 269
- first nonlinear version (TLE), 263–264
- flexible functional forms, 250–251
- generalized Leontief (GL), 267–268
- goodness of fit, 252
- log-quadratic model, flexibility
 estimation, cost function, 284
 parameter estimation, 284, 285
 parameterization, 283, 285
 quality of adjustment, 285–287
 sensitivity, 287
 unit cost, 287, 288
- multiplicative stochastic error, 253
- precision of fit, 252
- second nonlinear version (TLA)
 Euler's theorem, 264–265
 input-output ratio system, 266–267
 linear homogeneity property, 265–266
 own and cross price elasticities of demand, 267
 parameters, 265–266
 stochastic specification and technical change, 264
 unit cost formula, 266
- Sobolev norm, 251
- standard version
 Allen-Uzawa elasticities, 262
 analytical structure, 262
 CRS parameter, 261
 input cost share equations, 260
 nonhomotheticity and nonneutral technical change, 259
 output elasticity, cost function, 261
 rate of technical change, 261
- symmetric generalized McFadden (MF), 268–269
- theoretical background
 concavity property, 257
 consistency properties, cost function, 254, 255
 cost minimization, 255, 256
 differentiation, cost function, 255
 duality theory, 254, 255
 Engel aggregation, 258
 input substitutability, 256
 price elasticities of demand, 257
 proportional change, 257, 258
 rate of technical change, 258
- T4 Supplementary Tax File, 164
- V**
- Value-added based capital productivity index, 23
- Value-added based labour productivity index, 23
- Value-added based labour productivity level, 26
- Value-added based simple labour productivity index, 23
- Value-added based total factor productivity index, 22–23
- Voluntary separations
 empirical methodology, 164–167
 estimation, 171–172
 EUKLEMS data, 163–164
 industry shutdown rates, effects of, 163
 LWF, 162–164

probability of, 167–170
ways of, 162

W

Worker separations and potential job instability
empirical methodology, 164–167

estimation, 171–172
EUKLEMS data, 163–164
industry shutdown rates, effects of, 163
LWF, 162–164
probability of, 167–170
ways of, 162
World Productivity Database (WPD), 318