Springer Proceedings in Business and Economics

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

Productivity and Inequality

Springer

Springer Proceedings in Business and Economics

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

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

Productivity and Inequality

Springer

Editors
William H. Greene
Department of Economics
New York University
New York, NY, USA

Paul Makdissi
University of Ottawa
Ottawa, ON, Canada

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

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

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

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 978-3-319-68677-6 ISBN 978-3-319-68678-3 (eBook)
https://doi.org/10.1007/978-3-319-68678-3

Library of Congress Control Number: 2017963543

Printed on acid-free paper

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

Introduction

This book is the result of a selection of papers accepted at the North American Productivity Workshop (NAPW), which is a major biennial conference of researchers and practitioners on productivity and efficiency issues that is held every two years. NAPW IX was being held between June 15 and June 18, 2015, in Quebec City, and the theme of the conference was productivity and inequality. The 2016 conference included several keynote speeches by leading international productivity experts and presentations of specialized productivity-related research such as Susanto Basu, Professor of Economics (Boston College); Allan Collard-Wexler, Associate Professor of Economics (Duke University); Russell Davidson, Professor of Economics (McGill University); Erwin Diewert, Professor in the Vancouver School of Economics (University of British Columbia); Jonathan David Ostry, Deputy Director of the Research Department (IMF); Ariel Pakes, Thomas Professor of Economics (Harvard University); and Robin C. Sickles, Reginald Henry Hargrove Professor of Economics, Professor of Statistics (Rice University).

Each plenary session allowed for the presentation of recent and ongoing research in the areas of productivity, inequality, efficiency, data envelopment analysis, and index number theory. The plenary sessions were organized to cover a broad range of productivity, inequality, and efficiency topics. Basu looked at the general topic of productivity and the welfare of nations; Collard-Wexler showed how to estimate production functions with measurement error in inputs; Diewert presented a decomposition of US business sector TFP growth into technical progress and inefficiency components; Davidson presented the challenges of making statistical inference with income distributions; Ostry discussed redistribution, inequality, and growth; Pakes looked at new entries and new markets; and Sickles discussed the sources of income inequality by exploring the role of productivity growth.

The conference included many other researchers (150) from around the world (28 countries: Australia, Austria, Belgium, Brazil, Canada, China, Denmark, France, Germany, Greece, Hungary, Italy, Japan, Luxembourg, Macedonia, the Netherlands, Norway, Poland, Portugal, Romania, Russia, Spain, Sweden, Switzerland, Taiwan, Tunisia, the United Kingdom, and the United States), and the exposure of their research was a real gain for the Canadian community that works on issues of

productivity and inequality. The conference was hosted by the Department of Economics of Carleton University and organized in collaboration with Industry Canada, Network to Study Productivity in Canada from a Firm-Level Perspective, the Centre for Monetary and Financial Economics, and Bank of Canada.

The quality of the research papers presented at this conference attracted also representatives from Industry Canada, Bank of Canada, and Statistics Canada. Their participation was very important as the research ideas presented at the conference are very helpful for policy makers and can be used to formulate policies that can improve productivity performance of Canadian companies and, at the same time, can reduce inequality among Canadians.

Contents

Estimating Efficiency in the Presence of Extreme Outliers: A Logistic-Half Normal Stochastic Frontier Model with Application to Highway Maintenance Costs in England

Alexander D. Stead, Phill Wheat, and William H. Greene

Abstract In Stochastic Frontier Analysis the presence of outliers in the data, which can often be safely ignored in other forms of linear modelling, has potentially serious consequences in that it may lead to implausibly large variation in efficiency predictions when based on the conditional mean. This motivates the development of alternative stochastic frontier specifications which are appropriate when the two-sided error has heavy tails. Several existing proposals to this effect have proceeded by specifying thick tailed distributions for both error components in order to arrive at a closed form log-likelihood. In contrast, we use simulation-based methods to pair the canonical inefficiency distributions (in this example half-normal) with a logistically distributed noise term. We apply this model to estimate cost frontiers for highways authorities in England, and compare results obtained from the conventional normal-half normal stochastic frontier model. We show that the conditional mean yields less extreme inefficiency predictions for large residuals relative to the use of the normal distribution for noise.

Keywords Stochastic frontier · Normal · Logistic · Outliers · Maximum simulated likelihood

The original version of this chapter was revised. An erratum to this chapter can be found at https://doi.org/10.1007/978-3-319-68678-3_15

A.D. Stead • P. Wheat
Institute for Transport Studies, University of Leeds, Leeds, UK
e-mail: a.d.stead@leeds.ac.uk

W.H. Greene (✉)
Department of Economics, Stern School of Business, New York University, New York City, NY, USA
e-mail: wgreene@stern.nyu.edu

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_1

1 Introduction

The aim of frontier analysis is to estimate a frontier function based on efficient, or at least best-practice in sample, production and cost relationships against which the efficiency of firms and other decision making units (DMU) can be measured. A challenge for such analyses is dealing with the existence of noise, resulting from random shocks and measurement error in the dependent variable — in the data. In particular, in the presence of outliers, there can be a disproportionate impact on the estimated frontier and on all predictions of efficiency relative to it. The Data Envelopment Analysis (DEA) model (Charnes et al. 1978) and related mathematical programming approaches are deterministic, in that any noise present is attributed wholly to variation in efficiency, and are therefore particularly sensitive. This is also the case with some of the cruder econometric methods, such as Corrected Ordinary Least Squares (COLS). Here we focus instead on Stochastic Frontier Analysis (SFA) which should be more robust to noise given this is considered explicitly alongside inefficiency in the model formulation.

The specific motivation for this paper comes from an issue arising from the authors' work studying cost efficiency in a number of datasets. The example used in this paper is cost analysis of highways maintenance operations of local government authorities in England, which utilises bespoke data on operating and capital expenditure provided by each authority. When we compute the standard Jondrow et al. (1982) predictor, an implausibly wide range of efficiency scores is found. This issue is caused by large estimated error variances; in particular, a large VAR(u) will lead to a large spread of efficiency scores, while a large VAR(v) will lead to a greater degree of shrinkage of efficiency predictions toward the unconditional mean (Wang and Schmidt 2009). Large error variances are in our dataset caused by the presence of a relatively large number of outliers in the data, due to a combination of under- or over-reporting, unobserved investment cycle effects, and extreme weather events.

In this paper we consider methods to better deal with noise data in the stochastic frontier setting. We consider alternative methods which are better suited to handling outliers in the data, i.e. heavier tails in the error. After consideration of possible existing approaches, this leads us to propose a new stochastic frontier model with a logistic distribution for the noise error. This model is easy to estimate and has been programmed into a bespoke version of LIMDEP.

The structure of this paper is as follows: Section 2 reviews the received methods available to handle a large number of outliers in frontier analysis, and reviews the relevant literature and Sect. 3 introduces a logistic-half normal stochastic frontier (SF) models for dealing with heavy-tailed noise. Section 4 applies these models to our data on highways maintenance costs in England and compares the results to those obtained from the standard normal-half normal SF model, and Sect. 5 gives our summary and conclusions.

2 Literature Review: Potential Approaches to Dealing with Outliers

2.1 Adopting Alternative Predictors for Inefficiency

Before considering amendments to the standard stochastic frontier model, it is natural to ask whether there are alternative efficiency predictors which yield more intuitive distributions of predictions. Given that in cross sectional models, point predictors are known to be inconsistent for the quantity of interest; namely the firm specific realisation of a random variable (Wheat et al. 2014), then several point and interval predictors could be candidates.

One candidate is the conditional mode predictor (Jondrow et al. 1982) which, for the normal-half normal model, treats all observations with positive (negative) residuals in the production (cost) frontier case as fully efficient; likewise in the normal-exponential model, all residuals past a certain threshold—i.e. the inverse of the product of the squared rate parameter from the exponential component and the standard deviation of the normal component—are predicted to be fully efficient. The conditional mode predictor therefore yields more intuitive efficiency predictions at the top relative to the conditional mean. This is because the conditional mean for all firms will always be less than one (for VAR(u) > 0) and, in the case of large VAR(v) i.e. data with many outliers, this difference is likely to be non-trivial even for the best performing DMU (due to substantial shrinkage to the unconditional mean (Wang and Schmidt 2009)). Furthermore, for all other observations the conditional mode predictor yields a predicted efficiency score higher than that from the conditional mean predictor; the latter difference, however, tends to be small in magnitude at the bottom, and its usefulness in remedying implausibly low efficiency scores is therefore limited.

Another approach is to calculate prediction intervals, which show the range of plausible efficiency predictions for a given observation. Since in the normal-half normal case the conditional distribution of u is that of a truncated normal random variable (Jondrow et al. 1982), Horrace and Schmidt (1996) propose simply using the quantile function for this distribution to compute the upper bound of a prediction interval, which is also derived by Bera and Sharma (1999). However, Wheat et al. (2014) note that this method does not necessarily yield a minimum width interval, and derive minimum width intervals for the normal-half normal case, and discuss various methods of accounting for parameter uncertainty in computing prediction intervals. The use of prediction intervals in cases where predicted efficiency values are at the extremes could be useful in that they allow us to qualify our point predictions of efficiency by explicitly recognising that there are in fact a range of probable values which efficiency can take; however, this is not a solution to the underlying problem and of course, the range of probable values will include values even more implausible than the point predictor.

Overall, while alternative predictors are useful in SFA in general, the mass of the conditional distribution for the most efficient firm in our sample is still far from

zero (even if the peak of the distribution—i.e. the mode—is zero. Thus the question remains as to whether an alternative formulation of the stochastic frontier model could yield a more intuitive distribution of efficiency predictions. In particular a formulation which puts more weight on outlying observations being the result of noise rather than inefficiency seems to be appropriate. We now consider possible means to achieve this.

2.2 Heteroskedastic Stochastic Frontier Models

The basic SFA model assumes that both error components are homoskedastic, i.e. that they have a constant variance. Outliers in the data could result from heteroskedasticity in one or both error components, so that certain observations have a higher error variance than others. Discussion of heteroskedastic SF models have tended to focus on heteroskedasticity in the one-sided error; Reifschneider and Stevenson (1991) propose a normal-half normal model in which $\sigma_{ui} = g(U_i)$, $g(U_i) \in (0, \infty)$, Caudill and Ford (1993) propose a normal-half normal model in which $\sigma_{ui} = \sigma_u(U_i\gamma)^{\delta}$, and Caudill et al. (1995) propose a normal-half normal in which $\sigma_{ui} = \exp(U_i\gamma)$, where in each case U_i is a vector of explanatory variables including an intercept. Wang (2002) combined the Battese and Coelli (1995) specification of the pre-truncation mean of a truncated normal one-sided error in which $\mu_i = Z_i\beta$, where Z_i is again a vector of explanatory variables, with a slight variation in the Caudill et al. (1995) specification of the one-sided error variance so that $\sigma_{ui}^2 = \exp(U_i\gamma)$ into a single model, which has the additional advantage of allowing for non-monotonic relationships between inefficiency and explanatory variables.

In terms of handling outliers where these are assumed to reflect an unusually high variance in noise, it is more useful to allow for heteroskedasticity in the two-sided error, however; Wang and Schmidt (2009) show for the normal-half normal model that $E(u_i|\,\varepsilon_i)$ is a shrinkage of u_i towards $E(u_i)$, and that because of this, as $\sigma_{vi} \to 0, E(u_i|\,\varepsilon_i) \to u_i$, while as $\sigma_{vi} \to \infty, E(u_i|\,\varepsilon_i) \to E(u_i)$. Allowing for heteroskedasticity in v therefore allows for varying levels of shrinkage. Hadri (1999) introduces a doubly heteroskedastic SF model in which the variances of both error components are a function of vectors of explanatory variables U_i and V_i —which need not be the same—such that $\sigma_{ui} = \exp(U_i\gamma), \sigma_{vi} = \exp(V_i\theta)$. Finally, Kumbhakar and Sun (2013) introduce a normal-truncated normal model which combines the Battese and Coelli (1995) and Hadri (1999) specifications into a model in which the pre-truncation mean of the one-sided error, as well as the variances of both error components are functions of vectors of explanatory variables, so that $\mu_i = Z_i\beta$, $\sigma_{ui} = \exp(U_i\gamma)$, $\sigma_{vi} = \exp(V_i\theta)$.

Allowing for greater levels of variance in outlying observations is effectively another method of allowing for a heavy tailed distribution. The problem with adopting this approach using existing heteroskedastic SF models is that an appropriate variable is needed for inclusion in the variance function. A dummy variable identi-

fying outlying observations could be used, for example, however the identification of such outlying observations would either have to be done on an ex-post basis, or with reference to some arbitrary partial metric, and of course there is an added degree of arbitrariness in defining the cut-off point beyond which an observation is deemed to be outlying.

2.3 Thick Frontier Analysis

Berger and Humphrey (1991, 1992) introduced Thick Frontier Analysis (TFA), which is motivated by the observation of heavy-tailed errors in cost studies—specifically, in the banking sector—but in contrast to the present study assumes that this reflects a wide spread of efficiencies, rather than outliers in the data. In TFA, DMUs are sorted into quantiles based on some partial measure, e.g. unit cost, and separate regressions are run for the top and bottom quantiles. DMUs in the lowest and highest unit cost quantiles are implicitly judged to be equally efficient, with their residuals reflecting only error and luck. The difference in predicted unit costs for different size classes is then decomposed into exogenous market factors, i.e. that explained by differences in output mix, input prices, etc., and the remainder, which is regarded as inefficiency.

TFA has a number of disadvantages, such as the implicit assumption of equal efficiency among DMUs in the same quantile, and the implicit need for rather large sample sizes so that samples can be sensibly divided in this way. Also problematic is the arbitrariness of both the partial measure according to which DMUs are placed into quantiles, and the number of quantiles specified; Wagenvoort and Schure (1999) provide a solution to the latter problem, using a recursive algorithm by which, starting with OLS on the full sample of observations, the sample is divided into successively larger numbers of quantiles until the Lagrange multiplier test proposed by Breusch and Pagan (1980) fails to reject normality of the error term. However, the successive increases in the number of quantiles will require larger and larger sample sizes, and will tend to increase the distortionary effect of outlying observations on the estimated quantile regression lines, and hence on efficiency predictions.

The impact of outliers on efficiency scores in TFA is somewhat ambiguous. On one hand, the impact of outliers on efficiency scores will tend to be muted by the attribution of the residuals from the quantile regressions to noise, and by construction the DMUs in the top quantile will be judged fully efficient, while on the other hand the quantile regressions themselves will be more sensitive to outliers, which could lead to an exaggerated gap between the quartile regression lines, and hence an exaggerated range of inefficiency scores. This in fact reflects the different motivations and assumptions behind TFA, since as stated above, the underlying assumption behind TFA is that heavy tailed errors reflect a wide spread of inefficiency, i.e. a heavy tailed distribution of inefficiency, rather than a heavy tailed distribution of noise, making TFA inappropriate for the purpose of the current study; we therefore do not pursue TFA any further.

2.4 Non-Gaussian Stochastic Frontier Models

Another possible method of dealing with the impact of outliers in the data on efficiency scores is to directly alter the distributional assumptions of the basic SF model such that the noise component of the composed error, rather than being normally distributed, follows an alternative symmetric distribution with heavier tails.

One candidate for this is the Student's *t* distribution, a heavy-tailed distribution which approximates normality for finite sample sizes. Tancredi (2002) proposes a model in which the two-sided error is *t* distributed and the one-sided error follows a half *t* distribution—thus generalising the original normal-half normal of Aigner et al. (1977) to allow for heavier tails in both components of the composed error—and shows that as the residual approaches infinity, the conditional distribution of the one-sided error (conditional on the composed error realisation) is concentrated around zero in the normal-half normal model, and is completely flat in the *t*-half *t* model; thus in the former case, an observation with a large positive residual is judged to be close to the frontier with high probability, while in the latter case it is judged to be basically uninformative, making the model better at handling such outliers. Applying both models to the Christensen and Greene (1976) dataset on US electric utilities, the author shows that the *t*-half *t* performs better than the normal-half normal, and that allowing for heavy tails in this way increases the evidence for inefficiency in the model and overturns the Ritter and Simar (1994) finding that the basic SF model does not fit the data significantly better than OLS.

Nguyen (2010) introduces three additional non-Gaussian SF models, having two-sided and one-sided errors that respectively follow Laplace and exponential, Cauchy and half Cauchy, and Cauchy and truncated Cauchy distributions. These models are considered in a cross-section context, with application to the Christensen and Greene (1976) dataset, and Cauchy-half Cauchy balanced and unbalanced panel data models with time invariant inefficiency are also introduced, with application to the US banking dataset and to the WHO health sector dataset used in Greene (2004). The usefulness of some of the aforementioned models is limited by the unjustifiable assumptions made in order to simplify their derivation: the Laplace-exponential model assumes the variances of the two error components to be the same, as does the Cauchy-half Cauchy model for balanced panel data with respect to the variance of the two-sided error and the (pre-truncation) variance of the one-sided error; the latter model further assumes only two time periods. Nevertheless, both the cross-section and unbalanced panel Cauchy-half Cauchy models appear acceptable, and results from the latter are presented by Gupta and Nguyen (2010).

Horrace and Parmeter (forthcoming) discuss SFA with a Laplace-distributed two-sided error generally, and introduce a Laplace-truncated Laplace model; this is shown to reduce to a Laplace-exponential model when the pre-truncation mean of the one-sided error is less than zero, and to a Least Absolute Deviations (LAD) regression when the variance of the inefficiency term is zero. It is also shown that the conditional distribution of inefficiency is constant when the residual is zero,

so that all observations with positive residuals are given an identical efficiency score; as with the t-half t, the model therefore treats outlying observations as less informative. Results from Monte Carlo simulations suggest that the Laplace-exponential model performs better than the normal-exponential model when the error is miss-specified, and that it is more likely to produce non-zero estimates of the variance in inefficiency when OLS residuals display the wrong skew. The Laplace-truncated Laplace model is applied to estimate a cost frontier using the US airline data used in Greene (2012).

An analogous Bayesian approach to non-Gaussian SFA exists; Tchumtchoua and Dey (2007), estimate a *t*-half *t* Bayesian SF model, and Griffin and Steel (2007) briefly discuss how to estimate *t*-half normal, *t*-exponential, and *t*-gamma Bayesian SFA models using the WinBUGS software package.

To summarize, the non-Gaussian SF models are a potential way of dealing with the impact of outliers on the spread of efficiency predictions in SFA, given the different way the models treat outliers; they also have the advantage of being less arbitrary than simply excluding observations, or than the other methods discussed. A drawback of the existing models, however, is that in order to arrive at closed form expressions for their log-likelihoods, they also adopted alternative—i.e. thick tailed—distributions for u, which limits both the effectiveness of the models in reducing the impact of outliers on the range of efficiency predictions, and comparability with conventional SF models; we therefore prefer a model in which only v is drawn from a thick tailed distribution.

3 The Logistic-Half Normal Stochastic Frontier Model

3.1 Formulation and Estimation

In this paper, our motivation is to amend the conventional stochastic frontier model to accommodate data with large reporting errors. The work on non-Gaussian SF models discussed above motivates us to propose a further model which departs from the previous literature in that it amends the noise error term only and retains all of the conventional SF assumptions on the inefficiency error and the relationship between error components and regressors. This allows us to understand the extent to which alternative assumptions on the noise error term influence the efficiency predictions all other things equal.

In SFA, we have a composed error ε consisting of a symmetric noise component v and an inefficiency component u which is drawn from some one-sided distribution, such that

$$\varepsilon = v - su \tag{1}$$

Where s takes on a value of one for a production frontier and minus one for a cost frontier. In our case, we assume that v is drawn from a logistic distribution, and that v is from a half-normal distribution, such that

$$f(v) = \frac{\exp\left(\frac{v}{\sigma_v}\right)}{\sigma_v\left[1 + \exp\left(\frac{v}{\sigma_v}\right)\right]^2} \tag{2}$$

$$f(u) = \begin{cases} \frac{2}{\sigma_u}\phi\left(\frac{u}{\sigma_u}\right), & su > 0 \\ 0, & su \leq 0 \end{cases} \tag{3}$$

Where σ_v and σ_u are scale parameters. The joint density of ε and u is given by

$$f(u,\varepsilon) = \begin{cases} \frac{\exp\left(\frac{\varepsilon+su}{\sigma_v}\right)}{\sigma_v\left[1+\exp\left(\frac{\varepsilon+su}{\sigma_v}\right)\right]^2}\frac{2}{\sigma_u}\phi\left(\frac{u}{\sigma_u}\right), & su > 0 \\ 0, & su \leq 0 \end{cases} \tag{4}$$

And the marginal density of ε is given by the convolution

$$f(\varepsilon) = \int_0^\infty \frac{\exp\left(\frac{\varepsilon+su}{\sigma_v}\right)}{\sigma_v\left[1 + \exp\left(\frac{\varepsilon+su}{\sigma_v}\right)\right]^2}\frac{2}{\sigma_u}\phi\left(\frac{u}{\sigma_u}\right)du \tag{5}$$

Which is an integral with no closed form. It is therefore not possible to give an analytic expression for the log-likelihood function, and to proceed with maximum likelihood estimation. In such a case, maximum simulated likelihood techniques—see Train (2003) for an introduction to simulation-based methods—allow us to overcome this obstacle and estimate our model. The method followed here was first outlined in the context of the normal-gamma SF model by Greene (2003). We begin by noting that the integral in (5) is simply the expectation of $f(v)$ given that u is drawn from a half normal distribution

$$h(u) = E\left[f(v)|\, u \geq 0\right], \quad u \sim N\left[\mu, \sigma_u\right] \tag{6}$$

And thus we can form a simulated probability density function for ε by averaging over Q draws from a half normal distribution. The usual method of taking draws from a non-uniform distribution is to note that the cumulative density function of a random variable follows a uniform distribution, and thus by inverting the cumulative density function we can have the value of the random variable in terms of a uniformly distributed random variable; this inverse cumulative density function can therefore be used to transform draws from a uniform distribution into draws from any given distribution. Thus to generate draw number q from the half normal

distribution of our inefficiency term u we have

$$u_q = \sigma_u \Phi^{-1}\left(\frac{1}{2} + \frac{F_q}{2}\right) \tag{7}$$

Where F_q is draw number q from a uniform distribution. This leads us to the simulated probability density function for ε

$$\tilde{f}(\varepsilon) = \frac{1}{Q}\sum_{q=1}^{Q} \frac{\exp\left(\frac{\varepsilon + su_q}{\sigma_v}\right)}{\sigma_v\left[1 + \exp\left(\frac{\varepsilon + su_q}{\sigma_v}\right)\right]^2} \tag{8}$$

And, introducing subscripts for observation i, the simulated log-likelihood function is

$$\ln SL = -N \ln Q - N \ln \sigma_v + \sum_{i=1}^{N} \ln \sum_{q=1}^{Q} \frac{\exp\left(\frac{\varepsilon_i + su_{qi}}{\sigma_v}\right)}{\left[1 + \exp\left(\frac{\varepsilon_i + su_{qi}}{\sigma_v}\right)\right]^2} \tag{9}$$

Which may be maximised like any conventional log-likelihood function, provided we have our draws from the uniform distribution forming the u_{qi}s.

3.2 *Efficiency Predictions*

The conditional density of u given ε, is the ratio of the joint distribution of v and u and the density of ε

$$f(u|\varepsilon) = \frac{f(v) f(\mathrm{u})}{f(\varepsilon)} \tag{10}$$

Which, in the logistic-half normal case, gives

$$f(u|\varepsilon) = \begin{cases} \dfrac{\exp\left(\frac{\varepsilon+su}{\sigma_v}\right)/\left[1+\exp\left(\frac{\varepsilon+su}{\sigma_v}\right)\right]^2 \frac{2}{\sigma_u}\phi\left(\frac{u}{\sigma_u}\right)/(\sigma_v)}{\int_0^\infty \frac{\exp\left(\frac{\varepsilon+su}{\sigma_v}\right)}{\sigma_v\left[1+\exp\left(\frac{\varepsilon+su}{\sigma_v}\right)\right]^2}\frac{2}{\sigma_u}\phi\left(\frac{u}{\sigma_u}\right)du}, & su > 0 \\ 0, & su \le 0 \end{cases} \tag{11}$$

The Jondrow et al. (1982) and Battese and Coelli (1988) point predictors for efficiency are exp[–E(u| ε)] and E[exp(–u| ε)], respectively; these are derived by solving the integrals

$$E\left(u|\varepsilon\right) = \int_0^\infty uf\left(u|\varepsilon\right) du \tag{12}$$

$$E\left[\exp(-u)|\,\varepsilon\right] = \int_0^\infty \exp(-u) f\left(u|\varepsilon\right) du \tag{13}$$

Which, in the logistic-half normal case, gives

$$E\left(u|\varepsilon\right) = \frac{1}{f\left(\varepsilon\right)} \int_0^\infty \frac{u \exp\left(\frac{\varepsilon+su}{\sigma_v}\right)}{\sigma_v\left[1+\exp\left(\frac{\varepsilon+su}{\sigma_v}\right)\right]^2} \frac{2}{\sigma_u} \phi\left(\frac{u}{\sigma_u}\right) du \tag{14}$$

$$E\left[\exp(-u)|\,\varepsilon\right] = \frac{1}{f\left(\varepsilon\right)} \int_0^\infty \frac{\exp(-u)\exp\left(\frac{\varepsilon+su}{\sigma_v}\right)}{\sigma_v\left[1+\exp\left(\frac{\varepsilon+su}{\sigma_v}\right)\right]^2} \frac{2}{\sigma_u} \phi\left(\frac{u}{\sigma_u}\right) du \tag{15}$$

Both of which, again, contain integrals with no closed form solutions. Simulation is therefore required to generate these point predictions: we substitute $\widetilde{f}\left(\varepsilon\right)$ for $f(\varepsilon)$, and the remaining integrals are the expectation of u and $\exp(su)$ respectively multiplied by the probability density function of v, given that u is drawn from a half-normal distribution; this leads us to the simulated expectations

$$\tilde{E}\left(u|\varepsilon\right) = \frac{1}{\tilde{f}\left(\varepsilon\right)} \frac{1}{R} \sum_{r=1}^{R} \frac{u_r \exp\left(\frac{\varepsilon+su_r}{\sigma_v}\right)}{\sigma_v\left[1+\exp\left(\frac{\varepsilon+su_r}{\sigma_v}\right)\right]^2} \tag{16}$$

$$\tilde{E}\left[\exp(-u)|\,\varepsilon\right] = \frac{1}{\tilde{f}\left(\varepsilon\right)} \frac{1}{R} \sum_{r=1}^{R} \frac{\exp\left[\frac{\varepsilon+(s+\sigma_v)u_r}{\sigma_v}\right]}{\sigma_v\left[1+\exp\left(\frac{\varepsilon+su_r}{\sigma_v}\right)\right]^2} \tag{17}$$

Which we use to generate our point predictions of cost efficiency. Note that draws from the uniform distribution are also therefore needed to generate efficiency predictions following estimation of the model. In the notation above we distinguish between draws to approximate $f(\varepsilon)$ using q and the additional draws required to compute the further integral in (16) and (17) using r. This is to minimise any simulation bias.

4 Application to Highways Maintenance Costs in England

In this section, we apply the logistic-half normal SF model to a unique dataset on highway maintenance costs in England. Responsibility for maintaining roads in England is divided between Highways England—until 2015 the Highways Agency—a government-owned company responsible for maintenance of the trunk road network, and the county councils and unitary authorities which are responsible for maintenance of the non-trunk roads in their respective areas. In recent years, local authorities have been under increasing pressure to demonstrate efficient practice or efficiency improvements in areas such as highway maintenance, e.g. by undertaking benchmarking exercises with peers. This study uses data from the CQC Efficiency Network,[1] which is used to analyse the cost efficiency of local authorities' highway maintenance activities.

Previous econometric studies of road maintenance costs have tended to focus of the question of marginal costs of usage, and what these imply for road pricing, rather than on the relative cost efficiency of local authorities. Previous studies estimate cost functions using data on renewals and maintenance costs for motorways and canton roads in Switzerland (Schreyer et al. 2002), Austrian motorways (Sedlacek and Herry 2002), national—i.e. trunk—roads in Poland (Bak et al. 2006; Bak and Borkowski 2009), roads in Sweden (Haraldsson 2006; Jonsson and Haraldsson 2008), and German motorways (Link 2006, 2009) and federal roads (Link 2014). Much of this work is summarized by Link (2014), who estimates two cost models: one in which, as the author argues should be the case, the size of the road network maintained is used as the scale variable, and a second in which passenger car traffic and goods vehicle traffic are used as scale variables can be derived; the author apparently does not consider using both network size and traffic as outputs in a single model. The only study to look at efficiency in the context of highway maintenance is that of Fallah-Fini et al. (2009), which uses applies DEA to data for eight counties from the US state of Virginia, using road area and a set of quality measures as outputs, and maintenance expenditure, traffic and equivalent single axle loads as inputs, and a set of climate factors as non-discretionary variables.

We use an unbalanced panel consisting of data on the 70 local authorities from England that were members of the CQC efficiency network during 2014–15 and supplied cost data for at least one of the 5 years from 2009–10 to that year; this gives us a total of 327 observations. Cost data were supplied to the network by each authority individually according to definitions decided by a working group of network members, relating to operating expenditure and capital expenditure—both divided into direct and indirect categories—on carriageway maintenance only, i.e. excluding related activities such as winter service and footway maintenance, on the basis that they should be understandable and yield consistent submissions; we use the sum of these, total expenditure, as our dependent variable. Nevertheless,

[1] See http://www.nhtnetwork.org/cqc-efficiency-network/home/

preliminary analysis of the data reveals large differences in unit costs with a large number of extreme outliers in both direction, which are clearly subject to some kind of reporting error. As a result, standard SF models, as discussed in Sect. 1, yields a wide range of efficiency predictions, motivating the development of the model presented here.

In line with the previous literature, we use road length and traffic as output variables; road lengths are included as our measure of scale, while traffic—in terms of passenger kilometres—we divide by road length and include as a density variable. Detailed breakdowns of overall network length into urban and rural roads and also by classification, the different classifications being, in order of importance, A roads, B roads, classified unnumbered roads, and unclassified roads; we refer to the latter two as C and U roads, respectively. B, C and U roads are always maintained by local authorities, while A roads can be either trunk, and therefore the responsibility of Highways England, or non trunk, maintained by local authorities. The road length data we use include B, C, U and non trunk A roads; motorways, denoted by the letter M, and trunk A roads, are not included. Likewise, we use traffic data supplied directly by the Department for Transport (DfT) which relate only to local-authority maintained roads.

We separate overall network length into urban and rural road lengths, and further include the lengths relating to each classification as proportions of the overall network length. We also include road condition indicators for each road classification—also from DfT sources—and as input prices we include a measure of median hourly wages in civil engineering for each NUTS1 region from the Annual Survey of Hours and Earnings (ASHE) published by the Office for National Statistics (ONS) and a national index of materials prices in road construction from the Department for Business, Innovation and Skills (BIS). We employ a modified Cobb-Douglas functional form, in which we include second-order terms relating to urban and rural road length. The cost frontier we estimate is

$$\begin{aligned}\ln TOTEX = {} & \beta_0 + \beta_1 \ln URL + \beta_2 \ln RRL + \beta_3 \ln URL^2 + \beta_4 \ln RRL^2 \\ & + \beta_5 \ln URL \ln RRL + \beta_6 \ln TRAFFIC + \beta_7 RDCA + \beta_8 RDCBC + \beta_9 RDCU \\ & + \beta_{10} PROP_{UA} + \beta_{11} PROP_{UB} + \beta_{12} PROP_{UC} + \beta_{13} PROP_{UU} \\ & + \beta_{14} PROP_{RA} + \beta_{15} PROP_{RB} + \beta_{16} PROP_{RC} + \beta_{17} YEAR + \beta_{18} \ln WAGE \\ & + \beta_{19} \ln ROCOSM + \varepsilon \end{aligned} \tag{18}$$

Where *TOTEX* is total expenditure on carriageway maintenance, *URL* and *RRL* are the lengths of an authority's urban and rural road networks, respectively, *TRAFFIC* is a traffic density measure—i.e. traffic count divided by total road network length—and *RDCA*, *RDCBC* and *RDCU* are the proportions of A roads, B and C roads, and unclassified roads where maintenance should be considered, weighted by the shares of their respective road classifications in the total road network length. $PROP_{UA}$ through to $PROP_{RC}$ are urban A roads, urban B roads, etc. as proportions of the total network length, with the proportion of rural unclassified

Table 1 Outputs from the logistic-half normal and normal-half normal models

	Logistic-Half Normal			Normal-Half Normal		
	Estimate	s.e.	Sig	Estimate	s.e.	Sig
β_0	16.0631	0.0956	***	16.0350	0.14502	***
β_1 (ln *URL*)	0.13443	0.11162		0.12738	0.17112	
β_2 (ln *RRL*)	0.90841	0.11836	***	0.91675	0.17943	***
β_3 (ln URL^2)	0.23534	0.04447	***	0.24091	0.06291	***
β_4 (ln RRL^2)	0.08315	0.01057	***	0.08503	0.01586	***
β_5 (ln *URL* ln *RRL*)	−0.07189	0.02944	**	−0.08083	0.04421	*
β_6 (ln *TRAFFIC*)	0.37956	0.10259	***	0.41532	0.15442	***
β_7 (*RDCA*)	0.44014	0.09675	***	0.46356	0.14373	***
β_8 (*RDCBC*)	−0.07142	0.02682	***	−0.07057	0.03909	*
β_9 (*RDCU*)	−0.00397	0.00324		−0.00519	0.00529	
β_{10} ($PROP_{UA}$)	8.28742	1.9879	***	7.80954	3.24067	**
β_{11} ($PROP_{UB}$)	1.982	2.27009		0.66161	3.86852	
β_{12} ($PROP_{UC}$)	0.62504	1.21835		0.44784	2.05441	
β_{13} ($PROP_{UU}$)	1.10074	0.56802	*	1.09028	0.83493	
β_{14} ($PROP_{RA}$)	2.57286	1.08575	**	2.1196	1.57145	
β_{15} ($PROP_{RB}$)	2.40330	1.10305	**	2.67772	1.5444	*
β_{16} ($PROP_{RC}$)	1.11517	0.67064	*	0.98277	0.98812	
β_{17} (*YEAR*)	0.04055	0.01105	***	0.04457	0.01661	***
β_{18} (ln *WAGE*)	0.82267	0.23264	***	0.89086	0.34002	***
$(1-\beta_{18})$ (ln *ROCOSM*)[1]	0.17733	–	–	0.10914	–	–
σ_u	0.54321	0.02541	***	0.56798	0.01482	–
σ_v	0.16005	0.00745	***	0.27642	0.03015	–
Log Likelihood	−188.52			−189.14		

Statistical significance at the: * 10% level, ** 5% level, *** 1% level
Notes: (1) Parameter is equivalent to $1-\beta_{18}$ due to the imposition of linear homogeneity in input prices

roads omitted to avoid perfect multicollinearity. Finally, we include a time trend, *YEAR*, and two input prices: *WAGE*, a measure of regional gross hourly wages in civil engineering, and *ROCOSM*, a national index of materials prices for road construction. All variables are mean-centred, and linear homogeneity in input prices is imposed by dividing our cost and wage variables by our materials price index, which drops out of the model.

Table 1 shows the parameter estimates and associated standard errors and significance levels from the logistic-half normal model, and for comparison, the normal-half normal model, both estimated in LIMDEP. Following Greene (2003), we use Halton draws rather than pseudorandom number generator to obtain our draws from the uniform distribution; we use 1000 draws, and find that further increases or small reductions in the number of draws do not significantly affect our results.

We can see that both models yield similar estimates for each parameter, and that most of our variables are found to be statistically significant at the 10%, 5%, or 1% levels. To underline the similarities between the two models, we note that the correlation between the predicted residuals from each model is 0.9994 (rank correlation 0.9993). The log likelihood for the logistic-half normal model is higher than the corresponding value for the normal-half normal model indicating a superior fit.

The parameter estimates indicate constant to decreasing returns to scale at the sample average (the p-value for the null hypothesis of constant returns is 0.2396, so we fail to reject it), with increasing returns to scale for smaller authorities, and increasing returns to traffic density. It is also noticeable that the significance associated with each of the frontier parameters increases using the logistic-half normal model relative to the normal-half normal model. This is unsurprising, since the use of a thick-tailed noise distribution increases the robustness of our parameter estimates to outliers.

Also of interest here are the estimated error variances, and how these differ between the two models. The variance of u is given in both cases by

$$\text{VAR}(u) = \frac{\pi - 2}{\pi}\sigma_u^2 \tag{19}$$

While the variances of v in the logistic-half normal and normal-half normal models, respectively, are given by

$$\text{VAR}(v) = \frac{\pi^2}{3}\sigma_v^2 \tag{20}$$

$$\text{VAR}(v) = \sigma_v^2 \tag{21}$$

Table 2 shows VAR(u) and VAR(v) for both the logistic-half normal and normal-half normal models, along with total error variance, VAR(ε). We can see that neither the overall error variance, nor its individual components, differ substantially between the two models.

In spite of their similar error variances, however, we expect that the logistic-half normal model will result in a significantly narrower distribution of predicted efficiency scores, given the very different way that the two models handle outliers, as discussed in Sect. 3.2. Cost efficiency predictions from both models are generated using the Jondrow et al. (1982) conditional mean predictor, which is shown in (16) for the logistic-half normal case.

Table 2 Estimated error variances

	Logistic-Half Normal	Normal-half normal
VAR(u)	0.107225	0.117227
VAR(v)	0.084279	0.07641
VAR(ε)	0.191504	0.193637

Table 3 Summary of efficiency scores

	Logistic-Half Normal	Normal-half normal
Minimum	0.408882	0.225086
Mean	0.708911	0.659549
Median	0.724585	0.682412
Maximum	0.879474	0.918035
Range	0.470592	0.692949

Table 3 shows some summary statistics relating to the resulting efficiency predictions from both models. The correlation between the two sets of efficiency predictions is high, at 0.997. However, comparing the ranges of the two sets of predictions, we can see that, as expected, the logistic-half normal model results in a far narrower distribution of efficiency predictions. This is due mostly to a very marked difference in the minimum predicted efficiency score, which is far higher in the logistic-normal model, from which the mean and the median predictions are also higher, though the difference is progressively smaller in each case. The maximum prediction, however, is smaller in the logistic-half normal model than in the normal-half normal model due to the way the model handles outliers in either direction, though as discussed in Sect. 2.1, the maximum prediction from both models would have been one if we had used the conditional mode predictor.

Figure 1 gives a more detailed comparison, showing kernel density estimates for both sets of efficiency scores. In this, we can see a greater number of observations with low predicted efficiency scores from the normal-half normal model generally, and higher efficiency predictions generally more common in the logistic-half normal model; the latter being in spite of the fact that, due to the model's handling of outlying observations, the highest several efficiency scores are somewhat lower than those from the normal-half normal model. Our model therefore seems to result in an overall more intuitive distribution of efficiency predictions, with far fewer at the bottom of the range with only a relatively small impact on predictions at the top.

Figure 2 shows the relationship between efficiency predictions and corresponding residuals in both models. Given the similarity of the estimated frontier parameters, the ranges of the residuals across the two models are very similar, as are the estimated error variances, but the relationship between the residuals and the efficiency predictions are significantly different; in the normal-half normal model, the slope of the function diminishes for large positive or negative residuals, but in the logistic-half normal model, in addition to the slope being gentler overall, this is much more pronounced, with the function becoming almost flat — i.e. there being very little change in efficiency predictions — at either end of the range. This suggests that, in line with our discussion of the way that the model treats outlying observations, efficiency predictions do not approach zero or one for extreme values of the residuals.

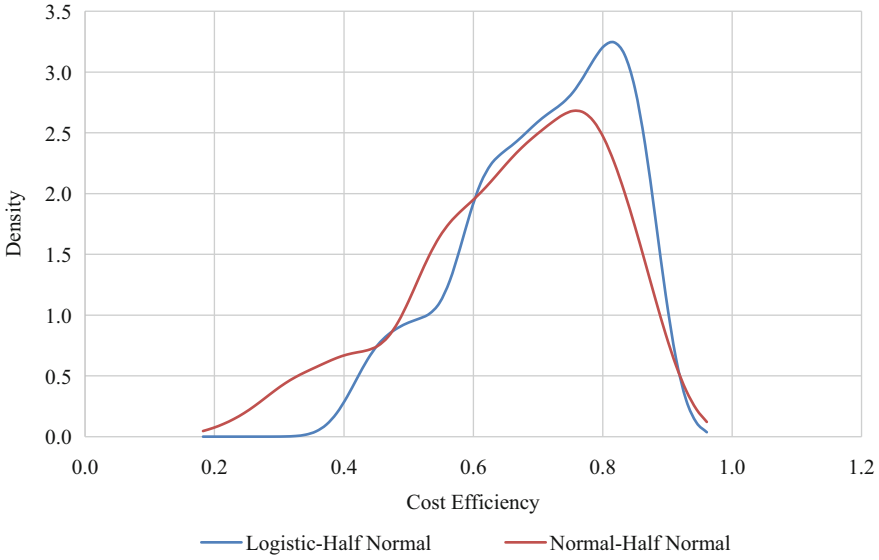


Fig. 1 Kernel densities of cost efficiency scores

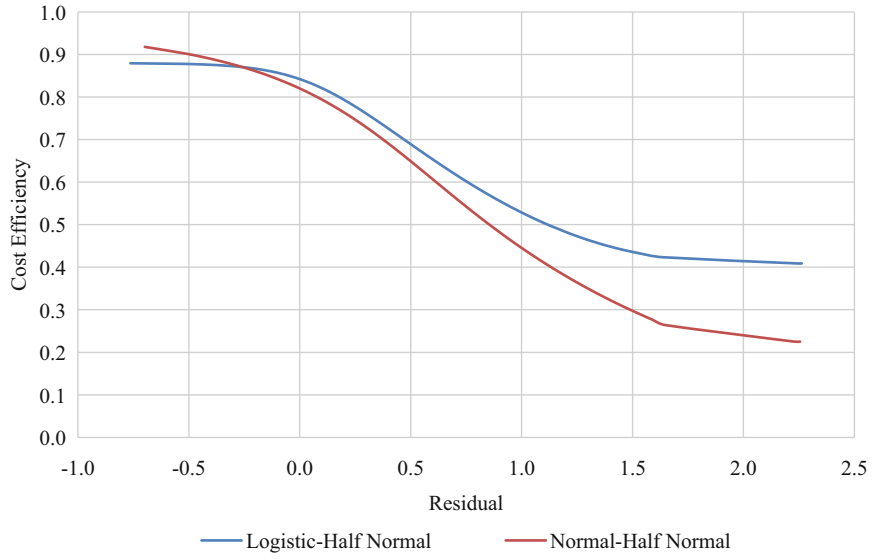


Fig. 2 Cost efficiency scores against residuals

5 Summary and Conclusions

This paper considers the issue of outliers and their impact on efficiency analyses. After reviewing how these issues have been handled in the existing literature, we have motivated and formulated a stochastic frontier (SF) model with a thick-tailed noise component. In contrast to previous models, in which both the noise and inefficiency terms have been drawn from a thick-tailed distribution, we use maximum simulated likelihood to estimate a model which combines a thick-tailed noise distribution—i.e. a logistic distribution—with a half normal inefficiency distribution. This model is easy to estimate and has been programmed into a bespoke version of LIMDEP. We show that the model handles outliers in both directions in a way that can produce a much narrower—and in the presence of outliers, more intuitive—range of efficiency predictions than standard SF models.

We apply our model to a unique dataset on highways maintenance costs in England, and compare the results to those from the normal-half normal SF model. The estimated frontier parameters and variances are found to be very similar to those from the normal-half normal model, but the former with greater significance due to the increased robustness of the model to outlying observations and we find, as expected, that the model results in a narrower range of efficiency predictions. The model is therefore effective in reducing the extent to which outlying observations are treated as having extreme efficiency values.

Further development could consider alternative distributions for u, such as truncated normal, exponential, or gamma, which would be easy to implement using our estimation approach. The issue of testing between our model and the standard SF model could also be explored. The authors are currently developing an alternative model in which v follows a Student's t distribution, which has the normal distribution as a limiting case, meaning that the model nests the standard SF model. A further advantage of the Student's t is that the thickness of the tails can be varied with its degrees of freedom parameter, making the model more general; a Student's t distribution with seven degrees of freedom is also a good approximation of the logistic distribution used in this study.

Acknowledgements The authors acknowledge funding from the CQC Efficiency Network (see http://www.nhtnetwork.org/cqc-efficiency-network/home/).

References

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

Bak, M., & Borkowski, P. (2009). *Marginal cost of road maintenance and renewal in Poland, CATRIN (Cost Allocation of TRansport INfrastructure cost) deliverable D6, annex 2*. Leeds: ITS, University of Leeds.

Bak, M., Borkowski, P., Musiatowicz-Podbial, G., & Link, H. (2006). *Marginal Infrastructure cost in Poland, marginal cost case studies for road and rail Transport deliverable D3, annex 1.2C*. Leeds: ITS, University of Leeds.

Battese, G. E., & Coelli, T. J. (1988). Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data. *Journal of Econometrics, 38*(3), 387–399.

Battese, G. E., & Coelli, T. J. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics, 20*(2), 325–332.

Bera, A. K., & Sharma, S. C. (1999). Estimating production uncertainty in stochastic frontier production function models. *Journal of Productivity Analysis, 12*(3), 187–210.

Berger, A. N., & Humphrey, D. B. (1991). The dominance of inefficiencies over scale and product mix economies in banking. *Journal of Monetary Economics, 28*(1), 117–148.

Berger, A. N., & Humphrey, D. B. (1992). Measurement and efficiency issues in commercial banking. In Z. Griliches (Ed.), *Output Measurement in the Service Sectors. NBER* (pp. 245–300).

Breusch, T. S., & Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies., 47*(1), 239–253.

Caudill, S. B., & Ford, J. M. (1993). Biases in frontier estimation due to heteroscedasticity. *Economics Letters, 41*(1), 17–20.

Caudill, S. B., Ford, J. M., & Gropper, D. M. (1995). Frontier estimation and firm-specific inefficiency measures in the presence of heteroscedasticity. *Journal of Business & Economic Statistics, 13*(1), 105–111.

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

Christensen, L. R., & Greene, W. H. (1976). Economies of scale in U.S. electric power generation. *Journal of Political Economy, 84*(4), 655–676.

Fallah-Fini, S., Triantis, K., & de la Garza, J. M. (2009). Performance measurement of highway maintenance operation using data envelopment analysis: Environmental considerations. In *IIE annual conference. Proceedings* (p. 693). Institute of Industrial Engineers-Publisher, Miami, USA.

Greene, W. H. (2003). Simulated likelihood estimation of the normal-gamma stochastic frontier function. *Journal of Productivity Analysis, 19*(2), 179–190.

Greene, W. H. (2004). Distinguishing between heterogeneity and inefficiency: Stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health Economics, 13*(10), 959–980.

Greene, W. H. (2012), *Econometric analysis* (7th ed.) Pearson.

Griffin, J. E., & Steel, M. F. J. (2007). Bayesian stochastic frontier analysis using WinBUGS. *Journal of Productivity Analysis, 27*(3), 163–176.

Gupta, A. K., & Nguyen, N. (2010). Stochastic frontier analysis with fat-tailed error models applied to WHO health data. *International Journal of Innovative Management, Information & Production., 1*(1), 43–48.

Hadri, K. (1999). Estimation of a doubly Heteroscedastic stochastic frontier cost function. *Journal of Business & Economic Statistics, 17*(3), 359–363.

Haraldsson, M. (2006). *Marginal cost for road maintenance and operation—A cost function approach, marginal cost studies for road and rail Transport deliverable D3, annex*. Leeds: ITS, University of Leeds.

Horrace, W. C., & Parmeter, C. F. (forthcoming). Forthcoming. A Laplace stochastic frontier model. *Econometric Reviews*.

Horrace, W. C., & Schmidt, P. (1996). Confidence statements for efficiency estimates from stochastic frontier models. *Journal of Productivity Analysis, 7*(2), 257–282.

Jondrow, J., Knox Lovell, C. A., Materov, I. S., & Schmidt, P. (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics, 19*(2), 233–238.

Jonsson, L., & Haraldsson, M. (2008). *Marginal costs of road maintenance in Sweden, CATRIN (cost allocation of TRansport INfrastructure cost) deliverable D6, annex 1*. Stockholm: VTI.

Kumbhakar, S. C., & Sun, K. (2013). Derivation of marginal effects of determinants of technical inefficiency. *Economics Letters, 120*(2), 249–253.

Link, H. (2006). An econometric analysis of motorway renewal costs in Germany. *Transportation Research Part A: Policy and Practice, 40*(1), 19–34.

Link, H. (2009). *Marginal costs of road maintenance in Germany, CATRIN (Cost Allocation of TRansport INfrastructure) deliverable D6, annex 3*. Stockholm: VTI.

Link, H. (2014). A cost function approach for measuring the marginal cost of road maintenance. *Journal of Transport Economics and Policy (JTEP), 48*(1), 15–33.

Nguyen, N. (2010). *Estimation of technical efficiency in stochastic frontier analysis*. PhD thesis, Bowling Green State University.

Reifschneider, D., & Stevenson, R. (1991). Systematic departures from the frontier: A framework for the analysis of firm inefficiency. *International Economic Review, 32*(3), 715–723.

Ritter, C., & Simar, L. (1994). *Another look at the American electrical utility data*. CORE Discussion Paper 9407. Centre for Operations Research and Econometrics, Catholic University of Louvain.

Schreyer, C., Schmidt, N., & Maibach, M. (2002). *Road econometrics–case study motorways Switzerland. UNITE (UNIfication of accounts and marginal costs for Transport efficiency) deliverable 10, annex A1b*. Leeds: ITS, University of Leeds.

Sedlacek, N., & Herry, M. (2002). *Infrastructure cost case studies. UNITE (UNIfication of accounts and marginal costs for Transport efficiency) deliverable 10, annex A1c*. Leeds: ITS, University of Leeds.

Tancredi, A. (2002). *Accounting for heavy tails in stochastic frontier models*. Working Paper No. 2002.16. Department of Statistical Sciences, University of Padua.

Tchumtchoua, S., & Dey, D. K. (2007). Bayesian estimation of stochastic frontier models with multivariate skew t error terms. *Communications in Statistics - Theory and Methods., 36*(5), 907–916.

Train, K. E. (2003). *Discrete choice methods with simulation* (2 (2009) ed.). Cambridge University Press, Cambridge, UK

Wagenvoort, J. L. M., & Schure, P. H. (1999). *The recursive thick frontier approach to estimating efficiency*. Report 99/02. European Investment Bank.

Wang, H.-J. (2002). Heteroscedasticity and non-monotonic efficiency effects of a stochastic frontier model. *Journal of Productivity Analysis, 18*(3), 241–253.

Wang, W. S., & Schmidt, P. (2009). On the distribution of estimated technical efficiency in stochastic frontier models. *Journal of Econometrics, 148*(1), 36–45.

Wheat, P., Greene, W., & Smith, A. (2014). Understanding prediction intervals for firm specific inefficiency scores from parametric stochastic frontier models. *Journal of Productivity Analysis, 42*(1), 55–65.

Alternative User Costs, Productivity and Inequality in US Business Sectors

W. Erwin Diewert and Kevin J. Fox

Abstract Using the new Bureau of Economic Analysis (BEA) Integrated Macroeconomic Accounts as well as other BEA data, we construct productivity accounts for two key sectors of the US economy: the Corporate Nonfinancial Sector (Sector 1) and the Noncorporate Nonfinancial Sector (Sector 2). Calculating user costs of capital based on, alternatively, ex post and predicted asset price inflation rates, we provide alternative estimates for capital services and Total Factor Productivity growth for the two sectors. Rates of return on assets employed are also reported for both sectors. In addition, we compare rates of return on assets employed and TFP growth rates when the land and inventory components are withdrawn from the asset base. Finally, implications for labour and capital shares from using alternative income concepts are explored.

Keywords User cost of capital · Total Factor Productivity · Rate of return on assets · Integrated Macroeconomic Accounts · Bureau of Economic Analysis · Ex post and ex ante asset inflation rates · US Nonfinancial Sector · Austrian model of production · Balancing rates of return · Inequality

JEL Classification: B25, C43, C82, D24, E22, E43

The first author, W. Erwin Diewert, gratefully acknowledges the financial support of the SSHRC of Canada, and both authors, W. Erwin Diewert and Kevin J. Fox, gratefully acknowledge the financial support of the Australian Research Council (LP0884095, DP150100830).

W.E. Diewert (✉)
Vancouver School of Economics, University of British Columbia, Vancouver, BC, Canada

The School of Economics, UNSW, Sydney, NSW, Australia
e-mail: erwin.diewert@ubc.ca

K.J. Fox
CAER & School of Economics, UNSW, Sydney, NSW, Australia
e-mail: K.Fox@unsw.edu.au

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_2

1 Introduction

The US Bureau of Economic Analysis (BEA), in conjunction with the Bureau of Labor Statistics (BLS) and the Board of Governors of the Federal Reserve, have developed a new set of production accounts, the Integrated Macroeconomic Accounts, for two major private sectors of the US economy: the Corporate Nonfinancial Sector (which we will call Sector 1) and the Noncorporate Nonfinancial Sector (which we will call Sector 2). For both sectors we work out the rate of return on assets employed back to 1960 and compute estimates of Total Factor Productivity (TFP) growth. In addition to comparing results across the sectors, we are particularly interested in determining whether rates of return and TFP growth have declined in recent years compared to the long run trends.

Another contribution is to document what can happen to user costs when ex post asset inflation rates are used in the user cost formula. Dale Jorgenson and his coworkers have advocated the use of ex post inflation rates in a user cost formula and so we call the resulting user costs "Jorgensonian". We show that for many assets, Jorgensonian user costs can be quite volatile and even negative at times which means that they cannot be used in many contexts. We advocate the use of predicted asset inflation rates in the user cost formula and we suggest a very simple moving average method for forming these predicted asset inflation rates, which we implement and compare with their Jorgensonian counterparts. We use Jorgensonian and predicted user costs to construct alternative measures of capital services and TFP growth for our two sectors of the US economy and, somewhat surprisingly, we find that there was little difference in the resulting trend measures of TFP growth, even though there are very large differences in the two sets of user costs.

An additional contribution is the examination of what happens to ex post rates of return on assets employed and on TFP growth as we withdraw assets from the asset base. This research has relevance for existing estimates of rates of return and TFP growth since many productivity studies exclude land and inventories from their asset base. We find that excluding these assets leads to exaggerated estimated rates of return on the remaining assets (as could be expected) but the effects on estimates of TFP growth are more variable. For our Sector 1, we found that excluding land and inventories had little effect on measured TFP growth but in Sector 2, the exclusion of land dramatically lowered measured TFP growth.

Finally, we use our data set to provide evidence on the debate regarding growing inequality due to a falling labour share in income. We find that moving from value added shares to (Hayekian) income shares provides stronger evidence of falling labour shares, indicative of growing inequality, for both our sectors.

Our accounting framework is laid out in the following section and the empirical results for the above measurement exercises follow in the subsequent sections.[1]

[1]The Appendix in Diewert and Fox (2016) explains in detail how we used the Integrated Macroeconomic Accounts to construct our data set for the two sectors of the US economy.

2 The Accounting Framework, User Costs and Rates of Return on Assets

Following Jorgenson and Griliches (1967), the Total Factor Productivity growth of a firm or industry of a sector is generally measured as an output index divided by an input index. The basic ingredients that go into an index number formula are two price vectors and two quantity vectors that list the output quantities and their prices (or the input quantities and their prices) produced or used for the production unit for the two observations being compared. Compiling prices and quantities for outputs and nondurable inputs for each period or observation is generally straightforward, but determining the flow price for a durable input is not straightforward. In order to accomplish the latter task, we will use a model of production that is due to the economist Hicks (1961) and the accountants Edwards and Bell (1961).[2] In each accounting period, the business unit combines the capital stocks and goods in process that it has inherited from the previous period with "flow" inputs purchased in the current period (such as labour, materials, services and additional durable inputs) to produce current period "flow" outputs as well as end of the period depreciated capital stock and inventory components which are regarded as outputs from the perspective of the current period (but will be regarded as inputs from the perspective of the next period). The model could be viewed as an Austrian model of production in honour of the Austrian economist Böhm-Bawerk (1891) who viewed production as an activity which used raw materials and labour to further process partly finished goods into finally demanded goods.[3] The beauty of this model is that a complex intertemporal production model with many periods can be reduced to a sequence of single period models.

Using this one period framework, we can now explain how user costs arise. Consider a production unit which produces quantities q_O of a single output, uses q_I units of an intermediate input, uses q_L units of labour services during say period t and purchases q_K units of a capital stock at the beginning of the period. After using the services of the capital input during period t, the production unit will have $q_K{}^u$ units of used (depreciated) capital on hand at the end of period t. We suppose that the production unit faces the positive prices $P_O{}^t$, $P_I{}^t$, $P_L{}^t$ for its output and variable inputs during period t and it faces the beginning of period t price for units of the capital input equal to $P_K{}^t$ and the price $P_K{}^{t+1u}$ for (used) units of the depreciated capital good at the end of period t. Finally, we assume that the production unit has a one period financial opportunity cost of capital at the beginning of period t (i.e., a beginning of the period nominal interest rate) equal to r^t. We also assume that the period t production possibilities set for this production unit is the set S^t.

[2]This model can be traced back in part to Walras (1954; 267–269) and Böhm-Bawerk (1891; 342) and more explicitly to von Neumann (1945; 2).

[3]For more on this Austrian model of production and additional references to the literature, see Diewert (1977; 108–111, 1980; 473, 2010, 2014a).

Using all of these assumptions, the production unit's (competitive) one period profit maximization problem is the following constrained optimization problem:

$$\max_{q_O, q_I, q_L, q_K, q_K^*} \left\{ P_O{}^t q_O - P_I{}^t q_I - P_L{}^t q_L - P_K{}^t q_K + \left(1 + r^t\right)^{-1} P_K{}^{t+1u} q_K{}^* \right.$$
$$\left. : (q_O, q_I, q_L, q_K, q_K{}^u) \in S^t \right\}. \tag{1}$$

Note that (1) assumes that all outputs and all variable inputs are paid for at the beginning of period t, as is the payment for the initial capital stock, which is an input. The depreciated capital stock $q_K{}^u$ is an output that is "produced" at the end of period t and its end of period t market value, $P_K{}^{t+1u} q_K{}^u$, is discounted by $(1 + r^t)$ to account for the opportunity cost of tying up financial capital in the asset over period t.

We make some additional assumptions at this point in order to further simplify the constrained optimization problem defined by (1). First we assume that the capital input depreciates at the constant geometric rate δ per period. The geometric model of depreciation has been advocated by Jorgenson (1989) and his coworkers and it is currently used by the BEA to construct US business sector capital stocks. The geometric model of depreciation implies that the depreciated quantity of end of period capital, $q_K{}^u$, is related to the corresponding beginning of the period capital stock, q_K, by the following equation[4]:

$$q_K{}^u = (1 - \delta)\, q_K \tag{2}$$

where δ is the geometric rate of depreciation and satisfies the inequalities $0 \leq \delta < 1$. Let $P_K{}^{t+1}$ be the end of period t price of a unit of the capital stock that has the same quality as the beginning of the period unit of the capital stock. Define the *constant quality asset inflation rate* over period t, i^t, by the following equation:

$$1 + i^t \equiv P_K{}^{t+1} / P_K{}^t. \tag{3}$$

Thus i^t is the constant quality inflation rate for the capital stock component from the beginning of period t to the end of period t. We assume that the anticipated end of period t price for the used beginning of the period capital stock is equal to the end of period price for a constant quality unit of the capital stock, i.e., we assume that $P_K{}^{t+1u} = P_K{}^{t+1}$ and thus we have the following equation:

$$P_K{}^{t+1u} = \left(1 + i^t\right) P_K{}^t. \tag{4}$$

[4]The assumption of Eq. (2) allows us to replace the initial production possibilities set S^t with a new set S^{t*} which is the feasible set of $(q_O{}^t, q_I, q_L, q_K)$.

Our final additional assumption is that all revenues and variable input costs are received and paid for at the end of period t instead of the beginning of period t. With these changes, the producer's constrained optimization problem becomes:

$$\max_{q_O,q_I,q_L,q_K} \left\{ \left(1+r^t\right)^{-1} \left[P_O{}^t q_O - P_I{}^t q_I - P_L{}^t q_L + \left(1+i^t\right)(1-\delta) P_K{}^t q_K\right] \right.$$
$$\left. - P_K{}^t q_K : (q_O, q_I, q_L, q_K) \in S^{t*} \right\}. \tag{5}$$

The terms involving q_K (the beginning of the period capital stock) in the objective function of (5) simplify to $-f_K{}^t \equiv -(1+r^t)^{-1}\left[1+r^t-(1+i^t)(1-\delta)\right]P_K{}^t$. Thus $f_K{}^t$ is the *discounted to the beginning of period t user cost of capital* using the geometric model of depreciation.[5] However, instead of discounting end of period cash flows to the beginning of period t, we could anti-discount or appreciate beginning of the period cash flows to the end of period t.[6] This can be accomplished by multiplying the objective function in (5) by $(1+r^t)$. If we do this, we obtain the following one period profit maximization problem:

$$\max_{q_O,q_I,q_L,q_K} \left\{ P_O{}^t q_O - P_I{}^t q_I - P_L{}^t q_L - u_K{}^t q_K : (q_O, q_I, q_L, q_K) \in S^{t*} \right\} \tag{6}$$

where the *end of period user cost of capital* $u_K{}^t$ is defined as follows:

$$u_K{}^t \equiv \left[1+r^t-\left(1+i^t\right)(1-\delta)\right]P_K{}^t = \left[r^t - i^t + \left(1+i^t\right)\delta\right]P_K{}^t. \tag{7}$$

This formula for the user cost of capital was obtained by Christensen and Jorgenson (1969; 302) for the geometric model of depreciation. It plays a fundamental role in our analysis.[7]

[5]This simple discrete time derivation of a user cost (as the net cost of purchasing the durable good at the beginning of the period and selling the depreciated good at an interest rate discounted price at the end of the accounting period) was developed by Diewert (1974; 504, 1980; 472–473, 1992; 194). Simplified user cost formulae (the relationship between the rental price of a durable input to its stock price) date back to Babbage (1835; 287) and to Walras (1954; 268–269). The original version of Walras in French was published in 1874. The early industrial engineer, Church (1901; 907–909) also developed a simplified user cost formula.

[6]Assuming that all of the flow transactions within the accounting period are realized at the end of each period is consistent with traditional accounting treatments of assets at the beginning and end of the accounting period and the cash flows that occur during the period; see Peasnell (1981; 56). The idea of anti-discounting to the end of the period to form *end of period user costs* $u_K{}^t$ (as opposed to the usual discounted to the *beginning of period user costs* $f_K{}^t$) was explicitly suggested by Diewert (2005a, b; 485). Anti-discounting is implicit in the derivation of the user cost of an asset using the geometric model of depreciation that was made by Christensen and Jorgenson (1969; 302).

[7]We have ignored tax complications in deriving (6). Any specific capital taxes (such as property taxes on real estate assets) should be added to the user cost formula for the relevant assets. In our empirical work, we were not able to obtain a breakdown of property taxes into land and

There are two versions of the user cost formula $u_K{}^t$ defined by (7) that we will use in this paper: (i) An *ex post version* that uses the actual beginning and end of period constant quality asset prices, $P_K{}^t$ and $P_K{}^{t+1}$, in order to define the asset inflation rate as $i^t \equiv (P_K{}^{t+1}/P_K{}^t) - 1$; and (ii) an *ex ante version* that uses the actual beginning of period t constant quality asset price, $P_K{}^t$, and an anticipated price for the asset at the end of period t, $P_K{}^{t+1*}$, in order to define an *anticipated asset inflation rate* as $i^* \equiv \left(P_K{}^{t+1*}/P_K{}^t\right) - 1$.

Jorgenson (1995, 1996) and his coworkers[8] have endorsed the use of ex post user costs, arguing that producers can perfectly anticipate future asset prices, and so we refer to the user costs defined by (7) when ex post asset inflation rates are used in the formula as *Jorgensonian user costs*. On the other hand, Diewert (1980; 476, 2005a; 492–493) and Hill and Hill (2003) endorsed the ex ante version for most purposes, since these ex ante user costs will tend to be smoother than their ex post counterparts and they will generally be closer to a rental or leasing price for the asset.[9] We will use our sectoral data on the US corporate and noncorporate financial sector to compute capital services aggregates and the resulting rates of TFP growth using both Jorgensonian and smoothed user costs that use predicted asset inflation rates.

We now discuss the issues surrounding the choice for the cost of capital, r^t, in the user cost formula. There are many methods for choosing r^t that have been suggested in the literature but the methods break down into two classes: those that choose exogenous estimates for r^t and those that choose r^t endogenously as the rate of return which will just make the value of inputs used during the period (including capital services) equal to the value of outputs produced during the accounting period. We will use endogenous estimates for the cost of capital in this study.[10]

In order to explain how the cost of capital is determined endogenously, we need to consider the case where the production unit uses N types of capital. Let $P_{Kn}{}^t$ and $P_{Kn}{}^{t+1}$ be the beginning and end of period t prices for a new asset of type n, let $0 \leq \delta_n < 1$ be the associated geometric depreciation rate, let $i_n{}^t \equiv (P_{Kn}{}^{t+1}/P_{Kn}{}^t) - 1$

structure components and so property tax rates are missing in our user costs that we construct in the following sections of this study. Business income taxes that fall on the gross return to the asset base can be absorbed into the cost of capital, r^t, so that r^t can be interpreted as the before income tax gross return to the asset base used by the production unit. For material on the construction of user costs for more complex systems of business income taxation, see Diewert (1992) and Jorgenson (1996).

[8] See in particular Jorgenson and Griliches (1967, 1972) and Christensen and Jorgenson (1969).

[9] Of course, the problem with using ex ante user costs is that there are many methods that could be used to predict asset inflation rates and these different methods could generate very different user costs. For empirical evidence on this point, see Harper, Berndt and Wood (1989), Diewert (2005a) and Schreyer (2012).

[10] The problem with the exogenous method is that it is difficult to determine exactly the appropriate external cost of financial capital. In particular, it is difficult to estimate the risk premium that is associated with investing in a production unit that generates variable ex post rates of return on its asset base over time. Nevertheless, the exogenous method is probably the preferred method from a theoretical point of view. These issues are discussed more fully in Schreyer, Diewert and Harrison (2005) and Schreyer (2009, 2012).

be the associated period t ex post asset n inflation rate over period t and let r^t be the endogenously determined period t ex post rate of return on the asset base for the production unit. The *ex post end of period t user cost for asset n* is defined as:

$$\begin{aligned} u_{Kn}{}^t &\equiv \left[1 + r^t - \left(1 + i_n{}^t\right)(1 - \delta_n)\right] P_{Kn}{}^t \\ &= \left[r^t - i_n{}^t + \left(1 + i_n{}^t\right)\delta_n\right] P_{Kn}{}^t; \qquad n = 1, \ldots, N. \end{aligned} \tag{8}$$

The period t technology set for the production unit is now the set of feasible production vectors $(q_O, q_I, q_L, q_{K1}, q_{K2}, \ldots, q_{KN})$ that belong to a period t production possibilities set S^{t*}. Let $q_O{}^t$, $q_I{}^t$, $q_L{}^t$ denote the period t output produced, intermediate input used and labour used for the production unit and let $(q_{K1}{}^t, q_{K2}{}^t, \ldots, q_{KN}{}^t)$ denote the vector of beginning of period t capital stocks used by the production unit during the period. The *ex post rate of return on the period t asset base*, r^t, is defined as the solution to the following (linear) equation which sets the value of period t outputs equal to the value of period t inputs where capital inputs are valued at their ex post user costs:

$$\begin{aligned} 0 &= P_O{}^t q_O{}^t - P_I{}^t q_I{}^t - P_L{}^t q_L{}^t - \Sigma_{n=1}{}^N u_{Kn}{}^t q_{Kn}{}^t \\ &= P_O{}^t q_O{}^t - P_I{}^t q_I{}^t - P_L{}^t q_L{}^t - \Sigma_{n=1}{}^N \left[1 + r^t - \left(1 + i_n{}^t\right)(1 - \delta_n)\right] P_{Kn}{}^t q_{Kn}{}^t. \end{aligned} \tag{9}$$

The ex post cost of capital method for determining the opportunity cost of capital that is based on solving Eq. (9) for r^t is due to Jorgenson and Griliches (1967, 1972) and Christensen and Jorgenson (1969). This method has been used frequently in the regulatory context. The method can be applied to both a single enterprise as well as to the economy as a whole. National statistical agencies that have programs that measure the productivity of market sector industries generally use this method.[11] From a national income accounting perspective, this method has the great advantage for statistical agencies that it preserves the structure of the *System of National Accounts 1993* SNA 1993 (Eurostat et al., 1993); i.e., the resulting user cost values just sum to the Gross Operating Surplus that was already in SNA 1993. Thus this method can be viewed as a straightforward elaboration of the present system of accounts which does not change its basic structure; it only provides a decomposition of Gross Operating Surplus or Cash Flow into more basic components.[12]

[11]The Bureau of Labor Statistics in the U.S. was the first to introduce an official program to measure Multifactor Productivity or Total Factor Productivity in 1983; see Dean and Harper (2001). Other countries with TFP programs now include Canada, Australia, the UK and New Zealand.

[12]This method for decomposing Gross Operating Surplus into explanatory factors (that are useful when measuring TFP growth), was endorsed in the *SNA 2008* (Eurostat et al., 2008); see Schreyer, Diewert and Harrison (2005) for a discussion of the issues.

In the following sections of this study, we will calculate these ex post rates of return on assets for our Sectors 1 and 2 and also use the Jorgensonian user costs defined by (8) when we calculate TFP growth rates for our two sectors.

The major disadvantage of using Jorgensonian user costs is their volatility and their tendency to become negative for at least some periods when asset inflation rates for particular assets (such as land) are high. These volatile and sometimes negative user costs do not approximate corresponding asset rental prices (when they exist), which do not exhibit the same volatility. Moreover, if these bouncing user costs are used in production function studies where the underlying technology is estimated using derived supply and demand functions, the resulting estimated parameters are unlikely to be reliable. Finally, if statistical agencies report these volatile user costs in their system of productivity accounts, users are likely to be skeptical of these estimates. Thus there is a need to produce smoother user costs for a variety of reasons.

Our approach to producing smoother user costs will be to use *predicted asset inflation rates*, say i_n^{t*}, in the user cost formula instead of the actual ex post asset inflation rates, i_n^t. The method for calculating these predicted asset inflation rates will be explained more fully in subsequent sections but the predicted rates are basically simple long run geometric averages of past ex post inflation rates. Once the smoothed or ex ante asset inflation rate for asset n in period t, i_n^{t*}, has been defined for n = 1,...,N, the *ex ante or smoothed end of period t user cost for asset n in period t*, u_{Kn}^{t*}, is defined as:

$$\begin{aligned} u_{Kn}^{t*} &\equiv \left[1 + r^{t*} - \left(1 + i_n^{t*}\right)(1 - \delta_n)\right] P_{Kn}^t \\ &= \left[r^{t*} - i_n^{t*} + \left(1 + i_n^{t*}\right)\delta_n\right] P_{Kn}^t; \qquad n = 1, \ldots, N \end{aligned} \tag{10}$$

where *the smoothed balancing rate of return for period t*, r^{t*}, is defined as the solution to the following equation (which is linear in r^{t*}):

$$\begin{aligned} 0 &= P_O^t q_O^t - P_I^t q_I^t - P_L^t q_L^t - \Sigma_{n=1}^N u_{Kn}^{t*} q_{Kn}^t \\ &= P_O^t q_O^t - P_I^t q_I^t - P_L^t q_L^t - \Sigma_{n=1}^N \left[1 + r^{t*} - \left(1 + i_n^{t*}\right)(1 - \delta_n)\right] P_{Kn}^t q_{Kn}^t. \end{aligned} \tag{11}$$

The smoothed rate of return r^{t*} can be viewed as a planned rate of return on assets that is expected on the beginning of the period value of the capital stock used by the production unit, provided expected asset inflation rates, the i_n^{t*}, are realized.[13] The

[13] Period t predicted prices for output, intermediate input and labour, say P_O^{t*}, P_I^{t*} and P_L^{t*}, should be used in equation (11) in order to calculate the period t predicted rate of return, r^{t*}, instead of the actual ex post prices for output, intermediate input and labour, P_O^t, P_I^t and P_L^t. However, it is the usual convention in production theory to assume that actual ex post unit value prices for variable outputs and inputs are equal to their predicted counterparts.

smoothed user costs defined by (10) will also provide a decomposition of Gross Operating Surplus into meaningful components. As we shall see, the ex ante user costs are considerably smoother than their Jorgensonian counterparts.[14] Note that both of our user cost models use endogenous rates of return. One of the main purposes of this study is to determine whether the choice of user cost formula affects our estimates of TFP growth.

We conclude this section by discussing some of the problems associated with the valuation of investments made by the production unit during period t and with the sales of assets that might have occurred during period t. We discuss these issues in the context of Eq. (9) but a similar discussion holds for the accounting framework defined by Eq. (11).

Consider the second equation in (9). Upon noting that $(1 + i_n{}^t)P_{Kn}{}^t$ is equal to the end of period t price of a new unit of the nth capital stock component, $P_{Kn}{}^{t+1}$, (9) can be rewritten as follows:

$$\begin{aligned} 0 = {} & P_O{}^t q_O{}^t - P_I{}^t q_I{}^t - P_L{}^t q_L{}^t - \Sigma_{n=1}{}^N \left(1 + r^t\right) P_{Kn}{}^t q_{Kn}{}^t \\ & + \Sigma_{n=1}{}^N P_{Kn}{}^{t+1} \left(1 - \delta_n\right) q_{Kn}{}^t. \end{aligned} \tag{12}$$

Recall our Austrian one period model of production where the beginning of period t capital stocks are regarded as inputs and the end of period capital stocks are regarded as outputs. The initial value of the capital stock, $\Sigma_{n=1}{}^N P_{Kn}{}^t q_{Kn}{}^t$, is appreciated to end of period values by multiplying this initial capital stock value by $(1 + r^t)$ so that the anti-discounted price for input asset n is $(1 + r^t)P_{Kn}{}^t$. Looking at (12), we see that the term $-\Sigma_{n=1}{}^N(1 + r^t)P_{Kn}{}^t q_{Kn}{}^t$ is (minus) the cost of the beginning of period t capital stock at end of period prices. The other prices on the right hand side of (12) are also expressed in end of period t prices. The first three terms on the right hand side of (12) correspond to the value of outputs produced during period t, less the value of intermediate and labour inputs used during the period. The final set of terms, $\Sigma_{n=1}{}^N P_{Kn}{}^{t+1}(1 - \delta_n)q_{Kn}{}^t$, is the end of period t value of the depreciated beginning of the period capital stock. Thus $(1 - \delta_n)q_{Kn}{}^t$ is the depreciated quantity of the beginning of the period capital stock for asset n that is left over at the end of period t. But this quantity is not the entire end of period t capital stock for asset n: during period t, there may have been investments in asset n. Suppose $q_{GIn}{}^t$ is the *gross investment* in asset n during period t (and

[14]There is a problem with interpreting these smoothed user costs as rental prices that might be anticipated at the beginning of the accounting period. When there is a severe recession in the economy in say period t, both r^t defined by solving (9) and r^{t*} defined by solving (11) will become unusually low (or even negative) and it is unlikely that the resulting low (or negative) user costs defined by (10) could be anticipated in practice. This limitation of our analysis should be kept in mind, particularly when looking at the user costs for 2008. This suggests that exogenous estimates for the cost of capital may be a more appropriate strategy for forming user costs that more closely approximate rental prices. If an exogenous r^{t*} is used, then equation (11) will not hold in general and it will be necessary to include pure profits (or losses) as a balancing item in the SNA. However, we do not pursue this line of inquiry in the present study.

the average price that the statistical agency assigns to this investment is $P_{GIn}{}^t$) for n = 1,...,N. Thus the actual end of period t quantity of asset n that the production unit has at its disposal is $q_{Kn}{}^{t+1} \equiv q_{GIn}{}^t + (1 - \delta)q_{Kn}{}^t$ and according to our accounting conventions, it should be valued at the end of period t asset price $P_{Kn}{}^{t+1}$. Hence the terms $\Sigma_{N=1}{}^N P_{Kn}{}^{t+1} q_{GIn}{}^t$ seem to be missing from the right hand side of (12). There is an explanation for this apparent puzzle.

Suppose asset n is a reproducible capital stock; i.e., an asset which is produced internally by the production unit or purchased from another producer. In this case, the value of the gross investment in asset n during period t, $P_{GIn}{}^t q_{GIn}{}^t$, will be part of the period t value of output for the production unit; i.e., it should be included as part of $P_O{}^t q_O{}^t$. This resolves the puzzle for reproducible capital stock components.[15]

Now suppose asset n is an inventory stock. External purchases of the inventory stock will be part of intermediate input purchases, $P_I{}^t q_I{}^t$. Sales of the inventory item will be reflected in the value of gross output, $P_O{}^t q_O{}^t$. But at the end of period t, there will be a net change in inventory stocks equal to $q_{Kn}{}^{t+1} - q_{Kn}{}^t$. Hence it appears that the term $P_{Kn}{}^{t+1}(q_{Kn}{}^{t+1} - q_{Kn}{}^t)$ is missing on the right hand side of (12). Note that since asset n is an inventory item, we assume $\delta_n \equiv 0$ and so the term $P_{Kn}{}^{t+1}(1 - \delta_n)q_{Kn}{}^t = P_{Kn}{}^{t+1} q_{Kn}{}^t$ is present on the right hand side of (12) and adding $P_{Kn}{}^{t+1}(q_{Kn}{}^{t+1} - q_{Kn}{}^t)$ to this term gives us the end of period value of inventory stocks, $P_{Kn}{}^{t+1} q_{Kn}{}^{t+1}$, which is the right answer from the perspective of the Austrian approach to production theory. But statistical agencies treat inventory change over a period as part of sectoral output and so the missing term $P_{Kn}{}^{t+1}(q_{Kn}{}^{t+1} - q_{Kn}{}^t)$ should be included as a part of the value of gross output, $P_O{}^t q_O{}^t$.[16] This resolves the puzzle for inventory components of the capital stock.

Suppose asset n is a type of land asset. As was the case for inventory items, we assume that the land depreciation rate is $\delta_n = 0$ and again, we find that the term $P_{Kn}{}^{t+1}(q_{Kn}{}^{t+1} - q_{Kn}{}^t)$ is missing on the right hand side of (12). This term now represents the value of net purchases of land of type n over period t, $q_{Kn}{}^{t+1} - q_{Kn}{}^t$, valued at end of period t price for this type of land, $P_{Kn}{}^{t+1}$. Statistical agencies typically do not treat land as an output or an intermediate input so in this case, the net quantity of land purchases over period t, valued at end of period land prices, will not appear as part of the gross output (if land was sold during period t) or intermediate input of the sector (if land was purchased during period t). Thus we need to treat

[15] However, to make the accounting precisely consistent with the Austrian model of production, we require that the price used to value gross investments in asset n during period, $P_{GIn}{}^t$, be equal to the end of period t imputed value for a unit of the nth capital stock. Setting $P_{Kn}{}^{t+1} = P_{GIn}{}^t$ will ensure consistency. In our empirical work, we used the BEA end of period price for reproducible units of the capital stock which may be slightly different from the corresponding investment price for the asset.

[16] The BEA in particular *does* include the value of inventory change as part of the gross output of an industry. However, they may not value the change in inventories at end of period prices of the inventory item and so again there may be a slight inconsistency in our empirical work due to this pricing difference. For a more complete treatment of the accounting problems associated with the treatment of inventories in the Austrian model of production, see Diewert (2005b).

these net purchases as an input cost item, so $-P_{Kn}{}^{t+1}(q_{Kn}{}^{t+1} - q_{Kn}{}^{t})$ should be added to the right hand side of (12), but this net cost value is offset by the increase in the value of land holdings at the end of the period, so $-P_{Kn}{}^{t+1}(q_{Kn}{}^{t+1} - q_{Kn}{}^{t})$ should be added to the right hand side of (12). These two entries cancel and so this resolves the puzzle for the land components of the capital stock.[17]

Real monetary balances are not regarded as productive inputs by national income accountants. However, we treat real monetary balances as being necessary for production.[18] Our accounting treatment of real balances is entirely analogous to our treatment of land and, as was the case with land, the accounting decomposition given by (9) or (12) is consistent with our Austrian theory of production.

Equations (9) and (12) provided an accounting treatment of production using ex post asset prices. As mentioned above, it is possible to build a similar accounting treatment of production using ex ante asset prices; i.e., instead of using Eq. (9) as our starting point for our accounting decomposition, we could have used Eq. (11). The consistency of Eq. (11) with the Austrian view of production is similar to our analysis of the consistency of Eqs. (9) and (12) with the Austrian approach to production theory.

We conclude this section with an important observation. Although we do not think that the Jorgensonian ex post user costs are useful in all contexts, we do think that *they are the right user costs to use in the context of finding the ex post rate of return on assets for a production unit*. Ex post rates of return are extremely important indicators of economic efficiency (along with TFP growth rates) and it is important to measure these rates of return accurately to guide the allocation of resources between sectors.[19]

Before we use the data that are described in the Appendix to construct ex post rates of return on assets and TFP growth rates, in the following section we describe the use of our data base to construct estimates for real wages and labour productivity.

[17] Suppose some land is purchased during period t at the price $P_{Kn}{}^{t^*}$ where this purchase price is not equal to the end of period price of land, $P_{Kn}{}^{t+1}$. The quantity of new land purchased will be equal to $q_{Kn}{}^{t+1} - q_{Kn}{}^{t}$. Then the term $-P_{Kn}{}^{t^*}(q_{Kn}{}^{t+1} - q_{Kn}{}^{t})$ should be added to the right hand side of (12) as a purchase of a primary input (a cost item) and at the same time, we should add the term $P_{Kn}{}^{t+1}(q_{Kn}{}^{t+1} - q_{Kn}{}^{t})$ to the right hand side of (12) to value this land purchase at the end of period t price of this type of land (a revenue item). Thus in principle, we should add the term $\left(P_{Kn}{}^{t+1} - P_{Kn}{}^{t^*}\right)(q_{Kn}{}^{t+1} - q_{Kn}{}^{t})$ to the right hand side of (12). If some land is sold during the period at the price $P_{Kn}{}^{t^*}$, then $q_{Kn}{}^{t+1} - q_{Kn}{}^{t}$ is negative and is equal to minus the quantity sold. In this case, we should still add the term $\left(P_{Kn}{}^{t+1} - P_{Kn}{}^{t^*}\right)(q_{Kn}{}^{t+1} - q_{Kn}{}^{t})$ to the right hand side of (12) to make the accounting consistent with our Austrian model of production. In our empirical work, we did not make these adjustments to the accounting identity given by (12); we simply assumed that $P_{Kn}{}^{t^*}$ is equal to our end of period price for the asset, $P_{Kn}{}^{t+1}$.

[18] This is consistent with the cash-in-advance, or vending machine model of the demand for money consider by Fischer (1974). For a more extensive discussion of the issues surrounding money in the production function, see Diewert and Fox (2015).

[19] See Harberger (1998) on the importance of the rate of return on assets.

3 Real Wages and Labour Productivity Growth in Sectors 1 and 2

In this section, we draw on our data base in order to calculate real wages and labour productivity for the two sectors.[20] We start with the data for Sector 1, the Nonfinancial Corporate Sector of the US economy.

Value Added of Sector 1 in year t, $V_{VA1}{}^t$ (in billions of dollars), and the corresponding year t price index, $P_{VA1}{}^t$ are listed in Table 1. Define the year t real value added of Sector 1 as $Q_{VA1}{}^t \equiv V_{VA1}{}^t/P_{VA1}{}^t$ for t = 1960,...,2014. The price and quantity of employee labour in Sector 1, are $P_{L1}{}^t$ and $Q_{L1}{}^t$ and define the value of labour input in Sector 1 for year t as $V_{L1}{}^t \equiv P_{L1}{}^t Q_{L1}{}^t$. The labour series $V_{L1}{}^t$ and $P_{L1}{}^t$ are also listed in Table 1. The value of capital services in Sector 1 for year t, $V_{KS1}{}^t$, can be defined residually by subtracting the value of labour input from value added; i.e., $V_{KS1}{}^t \equiv V_{VA1}{}^t - V_{L1}{}^t$. The shares of labour and capital services in value added are defined as $s_{L1}{}^t \equiv V_{L1}{}^t/V_{VA1}{}^t$ and $s_{KS1}{}^t \equiv V_{KS1}{}^t/V_{VA1}{}^t$. These Sector 1 value added shares along with the value of capital services are also listed in Table 1.

A beginning of year t price index for personal consumption expenditures, $P_C{}^t$, for t = 1960–2015, is converted to a centered consumer price index for year t, $P_C{}^{t^*}$, by averaging $P_C{}^t$ and $P_C{}^{t+1}$; i.e., define $P_C{}^{t^*} \equiv (1/2)\left(P_C{}^t + P_C{}^{t+1}\right)$ for t = 1960,...,2014.[21] This series, along with the wage rate index $P_{L1}{}^t$, was used to define the *Sector 1 real wage for year t*, defined as follows:

$$RW_1{}^t \equiv P_{L1}{}^t/P_C{}^{t^*}; \qquad t = 1960, \ldots, 2014. \tag{13}$$

Finally, *Sector 1 Labour Productivity in year t* (relative to the level in 1960), $ProdL_1{}^t$, is defined as follows (and is listed in Table 1):

$$ProdL_1{}^t \equiv \left[Q_{VA1}{}^t/Q_{L1}{}^t\right] / \left[Q_{VA1}{}^{1960}/Q_{VA1}{}^{1960}\right]; \qquad t = 1960, \ldots, 2014. \tag{14}$$

The price of (value added) output in Sector 1 grew 4.56 fold over the sample period while employee wages grew 13.95 fold. The geometric rates of growth were 3.61% per year for output and 5.00% per year for wages. Real wages grew 2.25 fold over the sample period while labour productivity grew 3.41 fold (the corresponding geometric rates of growth were 1.51% and 2.30% per year). The sample average labour and capital services shares were 68.6% and 31.4% respectively. The upward trend in the capital services share is noticeable in Fig. 1 which plots the series $s_{L1}{}^t$, $s_{KS1}{}^t$, $P_{VA1}{}^t$, $RW_1{}^t$ and $ProdL_1{}^t$. Note that the capital services share finishes up at 36.7%, well above its long term average of 31.4%. It can be seen that real wages have grown very slowly since 2007. Note also that real wage growth was fairly similar to labour productivity growth until 1982 and then labour productivity grew substantially faster than real wages. Finally, it can be seen that labour productivity

[20] This data base is described in more detail in the Appendix of Diewert and Fox (2016).

[21] This series was normalized to equal 1 in 1960. Note that the Sector 1 wage rate series $P_{L1}{}^t$ is also normalized to equal 1 in 1960.

Table 1 Sector 1 value added $V_{VA1}{}^t$, value of labour input $V_{L1}{}^t$, value of capital services $V_{KS1}{}^t$, value added shares of labour and capital services, $s_{L1}{}^t$ and $s_{KS1}{}^t$. price of labour $P_{L1}{}^t$, real wage $RW_1{}^t$ and labour productivity $ProdL_1{}^t$ for year t

Year	$V_{VA1}{}^t$	$V_{L1}{}^t$	$V_{KS1}{}^t$	$s_{L1}{}^t$	$s_{KS1}{}^t$	$P_{VA1}{}^t$	$P_{L1}{}^t$	$RW_1{}^t$	$ProdL_1{}^t$
1960	255.9	180.4	75.5	0.7050	0.2950	1.0000	1.0000	1.0000	1.0000
1961	262.8	184.5	78.3	0.7021	0.2979	1.0030	1.0290	1.0181	1.0301
1962	286.9	199.3	87.6	0.6947	0.3053	1.0095	1.0725	1.0505	1.0781
1963	306.1	210.1	96.0	0.6864	0.3136	1.0145	1.1094	1.0725	1.1231
1964	330.6	225.7	104.9	0.6827	0.3173	1.0239	1.1640	1.1104	1.1739
1965	364.7	245.4	119.3	0.6729	0.3271	1.0419	1.2085	1.1364	1.2152
1966	403.1	272.9	130.2	0.6770	0.3230	1.0723	1.2792	1.1753	1.2422
1967	423.9	291.1	132.8	0.6867	0.3133	1.0962	1.3501	1.2058	1.2643
1968	465.4	320.9	144.5	0.6895	0.3105	1.1302	1.4519	1.2537	1.3134
1969	504.4	356.1	148.3	0.7060	0.2940	1.1779	1.5545	1.2842	1.3178
1970	518.6	374.5	144.1	0.7221	0.2779	1.2216	1.6690	1.3173	1.3337
1971	558.5	396.2	162.3	0.7094	0.2906	1.2657	1.7739	1.3441	1.3927
1972	622.2	439.9	182.3	0.7070	0.2930	1.3108	1.8849	1.3791	1.4339
1973	698.7	495.1	203.6	0.7086	0.2914	1.3876	2.0157	1.4004	1.4452
1974	755.7	542.9	212.8	0.7184	0.2816	1.5239	2.2140	1.4058	1.4256
1975	818.1	569.0	249.1	0.6955	0.3045	1.6735	2.4284	1.4139	1.4708
1976	928.1	640.0	288.1	0.6896	0.3104	1.7549	2.6233	1.4415	1.5282
1977	1052.9	723.3	329.6	0.6870	0.3130	1.8543	2.8320	1.4697	1.5673
1978	1201.4	829.5	371.9	0.6904	0.3096	1.9868	3.0722	1.4886	1.5788
1979	1341.6	942.4	399.2	0.7024	0.2976	2.1497	3.3668	1.5002	1.5718
1980	1452.7	1030.7	422.0	0.7095	0.2905	2.3508	3.7310	1.5079	1.5770
1981	1641.7	1139.8	501.9	0.6943	0.3057	2.5530	4.0746	1.5100	1.6206
1982	1701.4	1183.3	518.1	0.6955	0.3045	2.7049	4.3920	1.5317	1.6459
1983	1817.5	1250.1	567.4	0.6878	0.3122	2.7546	4.5656	1.5253	1.6988
1984	2040.5	1388.2	652.3	0.6803	0.3197	2.8397	4.7838	1.5419	1.7456
1985	2172.9	1490.1	682.8	0.6858	0.3142	2.8900	5.0489	1.5719	1.7959
1986	2260.7	1578.2	682.5	0.6981	0.3019	2.9303	5.3247	1.6154	1.8349
1987	2425.7	1685.5	740.2	0.6949	0.3051	2.9859	5.5326	1.6337	1.8799
1988	2640.7	1825.3	815.4	0.6912	0.3088	3.0624	5.8270	1.6549	1.9406
1989	2772.6	1934.8	837.8	0.6978	0.3022	3.1554	6.0175	1.6431	1.9265
1990	2897.7	2037.5	860.2	0.7031	0.2969	3.2507	6.3362	1.6568	1.9542
1991	2946.1	2071.1	875.0	0.7030	0.2970	3.3222	6.6568	1.6788	2.0093
1992	3074.6	2188.7	885.9	0.7119	0.2881	3.3645	6.9976	1.7209	2.0597
1993	3216.0	2271.0	945.0	0.7062	0.2938	3.4346	7.1064	1.6915	2.0656
1994	3465.8	2398.7	1067.1	0.6921	0.3079	3.4868	7.2530	1.6773	2.1188
1995	3682.7	2524.6	1158.1	0.6855	0.3145	3.5344	7.3930	1.6759	2.1510
1996	3924.4	2667.7	1256.7	0.6798	0.3202	3.5578	7.6890	1.7111	2.2413
1997	4219.5	2862.6	1356.9	0.6784	0.3216	3.5860	7.9603	1.7545	2.3067
1998	4470.8	3093.8	1377.0	0.6920	0.3080	3.5955	8.4875	1.8610	2.4048
1999	4745.3	3310.0	1435.3	0.6975	0.3025	3.6193	8.8505	1.9151	2.4714

(continued)

Table 1 (continued)

Year	V_{VA1}^t	V_{L1}^t	V_{KS1}^t	s_{L1}^t	s_{KS1}^t	P_{VA1}^t	P_{L1}^t	RW_1^t	$ProdL_1^t$
2000	5063.1	3597.3	1465.8	0.7105	0.2895	3.6610	9.4524	2.0011	2.5618
2001	5026.2	3584.6	1441.6	0.7132	0.2868	3.7132	9.8280	2.0421	2.6163
2002	5066.0	3542.0	1524.0	0.6992	0.3008	3.7109	10.0373	2.0525	2.7272
2003	5228.7	3595.7	1633.0	0.6877	0.3123	3.7486	10.4012	2.0878	2.8444
2004	5577.0	3762.8	1814.2	0.6747	0.3253	3.8262	10.8273	2.1235	2.9567
2005	5958.9	3930.3	2028.6	0.6596	0.3404	3.9577	11.2147	2.1357	3.0287
2006	6377.9	4129.3	2248.6	0.6474	0.3526	4.0789	11.6165	2.1598	3.1010
2007	6571.4	4305.3	2266.1	0.6552	0.3448	4.1610	12.0778	2.1894	3.1233
2008	6624.1	4358.0	2266.1	0.6579	0.3421	4.2492	12.4226	2.1993	3.1326
2009	6253.9	4088.4	2165.5	0.6537	0.3463	4.3182	12.6545	2.2109	3.1601
2010	6605.7	4158.7	2447.0	0.6296	0.3704	4.3216	12.8807	2.2228	3.3375
2011	6921.7	4363.4	2558.3	0.6304	0.3696	4.4176	13.1669	2.2282	3.3331
2012	7321.5	4593.3	2728.2	0.6274	0.3726	4.4916	13.4816	2.2322	3.3727
2013	7591.9	4747.4	2844.5	0.6253	0.3747	4.5205	13.6248	2.2228	3.3978
2014	7895.8	4995.8	2900.0	0.6327	0.3673	4.5568	13.9548	2.2503	3.4121

Note: All values are in billions of dollars

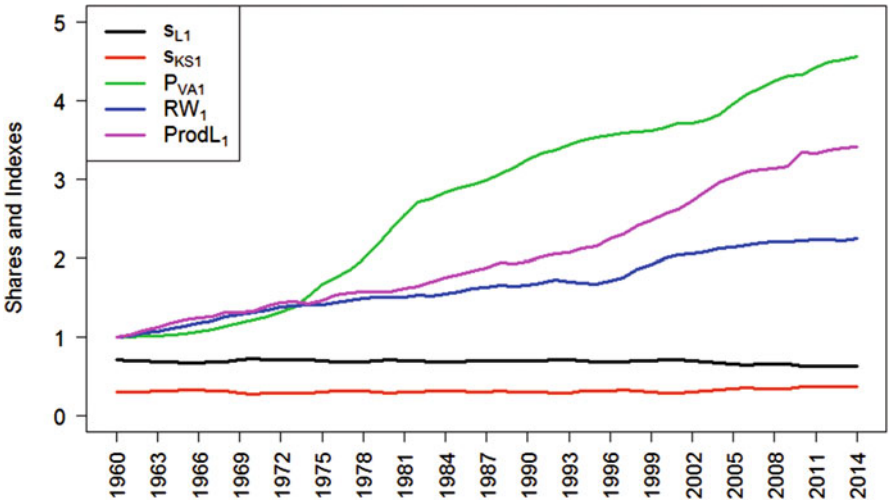


Fig. 1 Sector 1 labour and capital shares of value added, output price, real wage and labour productivity

in Sector 1 is still growing fairly steadily since 2006 at the geometric average rate of 1.20% per year but this rate is lower than the historical average rate of 1.51% per year.

We turn our attention to Sector 2, the Nonfinancial Noncorporate Sector of the US economy. Value Added of Sector 2 in year t, V_{VA2}^t (in billions of dollars) and the corresponding year t price index, P_{VA2}^t are listed in Table 2. Define the year t real value added of Sector 2 as $Q_{VA2}^t \equiv V_{VA2}^t/P_{VA2}^t$ for t = 1960,...,2014. The value and price of labour in Sector 2, V_{L2}^t and P_{L2}^t are listed in Table 2. The value of capital services in Sector 2 for year t, V_{KS2}^t, can be defined residually by subtracting the value of labour input from value added; i.e., $V_{KS2}^t \equiv V_{VA2}^t - V_{L2}^t$. The shares of labour and capital services in value added for Sector 2 are defined as $s_{L2}^t \equiv V_{L2}^t/V_{VA2}^t$ and $s_{KS2}^t \equiv V_{KS2}^t/V_{VA2}^t$. These Sector 2 value added shares along with the value of capital services are also listed in Table 2.

Again, we use the consumption price series $P_C^{t^*}$ along with the Sector 2 wage rate index P_{L2}^t to define the *Sector 2 real wage for year t*, $RW_2^t \equiv P_{L2}^t/P_C^{t^*}$ for t = 1960,...,2014. This series also appears in Table 2.

Finally, *Sector 2 Labour Productivity of Sector 1 in year t* (relative to the level in 1960), $ProdL_2^t$, is defined as follows (and listed in Table 2):

$$ProdL_2^t \equiv \left[Q_{VA2}^t/Q_{L2}^t\right] / \left[Q_{VA2}^{1960}/Q_{L2}^{1960}\right]; \quad t = 1960, \ldots 2014. \tag{15}$$

The price of (value added) output in Sector 2 grew 7.71 fold over the sample period (much higher than the Sector 1 price growth of 4.56 fold) while wages grew 12.67 fold. The geometric rates of growth were 3.61% per year for real value added, 3.86% per year for the value added deflator and 4.81% per year for wages. Real wages grew 2.04 fold over the sample period while labour productivity grew 2.36 fold, much lower than the 3.41 fold of labour productivity in Sector 1. The long run average geometric rates of growth of real wages and labour productivity for Sector 2 were 1.33% and 1.61% per year while the corresponding growth rates for Sector 1 were 1.51% and 2.30% per year. Thus real wage growth and labour productivity growth in Sector 2 were substantially below their Sector 1 counterparts. The sample average labour and capital services shares in Sector 2 were 56.7% and 43.3% (68.6% and 31.4% in Sector 1). It can be seen that Sector 2 is much more capital intensive than Sector 1. The upward trend in the capital services share is very noticeable in Fig. 2, which plots the series s_{L2}^t, s_{KS2}^t, P_{VA2}^t, RW_2^t and $ProdL_{L2}^t$. Note that the capital services share finishes up at 50.4%, well above its long term average of 43.3%. It can be seen that real wages have grown very slowly since 2001. Note also that real wage growth stagnated after 2007 while labour productivity continued to grow. *It can be seen that the structure of production is entirely different in the noncorporate nonfinancial sector as compared to the corporate nonfinancial sector*.

In the following section, we will calculate price and quantity indexes for the capital stocks used in both sectors as well as the corresponding real and nominal capital output ratios for the two sectors.

Table 2 Sector 2 value added V_{VA2}^t, value of labour input V_{L2}^t, value of capital services V_{KS2}^t, value added shares of labour and capital services, s_{L2}^t and s_{KS2}^t. price of labour P_{L2}^t, real wage RW_2^t and labour productivity $ProdL_2^t$ for year t

Year	V_{VA2}^t	V_{L2}^t	V_{KS2}^t	s_{L2}^t	s_{KS2}^t	P_{VA2t}	P_{L2}^t	RW_2^t	$ProdL_2^t$
1960	107.4	76.6	30.8	0.7135	0.2865	1.0000	1.0000	1.0000	1.0000
1961	110.2	76.8	33.4	0.6967	0.3033	1.0161	1.0297	1.0187	1.0378
1962	114.2	78.4	35.8	0.6864	0.3136	1.0306	1.0720	1.0500	1.0811
1963	117.1	79.2	37.9	0.6767	0.3233	1.0424	1.1087	1.0719	1.1216
1964	123.0	82.9	40.1	0.6737	0.3263	1.0599	1.1635	1.1100	1.1626
1965	130.0	84.7	45.3	0.6512	0.3488	1.0810	1.2053	1.1334	1.2216
1966	138.5	87.5	51.0	0.6318	0.3682	1.1142	1.2675	1.1646	1.2848
1967	142.1	89.4	52.7	0.6291	0.3709	1.1494	1.3281	1.1862	1.3104
1968	149.8	93.5	56.3	0.6241	0.3759	1.2014	1.4232	1.2289	1.3543
1969	157.8	99.1	58.7	0.6279	0.3721	1.2540	1.5223	1.2577	1.3795
1970	163.4	102.8	60.6	0.6294	0.3706	1.3032	1.6246	1.2823	1.4131
1971	173.1	106.8	66.3	0.6172	0.3828	1.3646	1.7248	1.3069	1.4613
1972	191.2	113.2	78.0	0.5919	0.4081	1.4302	1.8305	1.3393	1.5429
1973	223.5	124.1	99.4	0.5553	0.4447	1.4928	1.9487	1.3539	1.6773
1974	235.2	135.6	99.6	0.5763	0.4237	1.6334	2.1193	1.3456	1.6064
1975	252.0	144.1	107.9	0.5718	0.4282	1.8018	2.3053	1.3423	1.5966
1976	275.6	155.4	120.2	0.5638	0.4362	1.9586	2.4811	1.3634	1.6031
1977	300.7	170.1	130.6	0.5655	0.4345	2.1231	2.6649	1.3830	1.5835
1978	340.7	189.5	151.2	0.5562	0.4438	2.2748	2.8772	1.3941	1.6224
1979	380.3	211.6	168.7	0.5565	0.4435	2.5432	3.1382	1.3983	1.5821
1980	399.8	233.2	166.6	0.5832	0.4168	2.7131	3.4595	1.3982	1.5599
1981	435.5	253.4	182.1	0.5819	0.4181	2.9761	3.7703	1.3972	1.5534
1982	454.5	272.6	181.9	0.5998	0.4002	3.1681	4.0618	1.4166	1.5252
1983	480.7	290.0	190.7	0.6032	0.3968	3.4070	4.2399	1.4165	1.4720
1984	556.3	314.0	242.3	0.5644	0.4356	3.4506	4.4382	1.4305	1.6259
1985	600.4	330.0	270.4	0.5497	0.4503	3.6182	4.6619	1.4514	1.6725
1986	636.3	346.1	290.2	0.5439	0.4561	3.6821	4.8851	1.4821	1.7404
1987	667.5	365.3	302.2	0.5472	0.4528	3.8627	5.0870	1.5021	1.7171
1988	727.1	390.9	336.2	0.5376	0.4624	4.0499	5.3541	1.5206	1.7546
1989	774.1	413.1	361.0	0.5337	0.4663	4.2668	5.5205	1.5074	1.7298
1990	807.5	437.7	369.8	0.5420	0.4580	4.4449	5.7961	1.5155	1.7165
1991	815.0	458.3	356.7	0.5623	0.4377	4.6295	6.0487	1.5254	1.6579
1992	869.8	471.9	397.9	0.5425	0.4575	4.7022	6.3684	1.5661	1.7813
1993	903.5	501.4	402.1	0.5550	0.4450	4.8129	6.4767	1.5417	1.7301
1994	951.2	522.6	428.6	0.5494	0.4506	4.8685	6.6155	1.5299	1.7648
1995	992.0	541.2	450.8	0.5456	0.4544	5.0990	6.7770	1.5363	1.7381
1996	1069.5	566.7	502.8	0.5299	0.4701	5.3611	7.0498	1.5688	1.7707
1997	1136.9	601.8	535.1	0.5294	0.4706	5.5402	7.3528	1.6206	1.7889
1998	1229.4	637.3	592.1	0.5184	0.4816	5.6619	7.7929	1.7087	1.8943
1999	1312.3	662.7	649.6	0.5050	0.4950	5.7664	8.1347	1.7602	1.9933

(continued)

Table 2 (continued)

Year	V_{VA2}^t	V_{L2}^t	V_{KS2}^t	s_{L2}^t	s_{KS2}^t	P_{VA2t}	P_{L2}^t	RW_2^t	$ProdL_2^t$
2000	1420.7	712.9	707.8	0.5018	0.4982	6.0680	8.6602	1.8334	2.0293
2001	1637.1	830.6	806.5	0.5073	0.4927	6.3365	8.9251	1.8545	1.9809
2002	1707.0	861.7	845.3	0.5048	0.4952	6.3650	9.0834	1.8575	2.0170
2003	1800.5	934.1	866.4	0.5188	0.4812	6.4184	9.3894	1.8847	2.0118
2004	1953.7	1023.3	930.4	0.5238	0.4762	6.5464	9.7942	1.9209	2.0381
2005	2088.6	1102.8	985.8	0.5280	0.4720	6.6842	10.1199	1.9272	2.0460
2006	2293.1	1210.0	1083.1	0.5277	0.4723	6.8051	10.5245	1.9567	2.0912
2007	2356.3	1303.0	1053.3	0.5530	0.4470	6.9719	10.9629	1.9872	2.0289
2008	2474.5	1315.8	1158.7	0.5317	0.4683	6.9994	11.2585	1.9932	2.1583
2009	2321.0	1271.1	1049.9	0.5477	0.4523	6.8638	11.4066	1.9929	2.1651
2010	2395.5	1291.9	1103.6	0.5393	0.4607	7.0960	11.6811	2.0158	2.1779
2011	2592.9	1324.2	1268.7	0.5107	0.4893	7.2228	11.9747	2.0264	2.3162
2012	2742.3	1386.7	1355.6	0.5057	0.4943	7.3719	12.2744	2.0323	2.3493
2013	2839.8	1410.8	1429.0	0.4968	0.5032	7.5334	12.3977	2.0226	2.3635
2014	2966.3	1470.3	1496.0	0.4957	0.5043	7.7110	12.6666	2.0426	2.3645

Note: All values are in billions of dollars

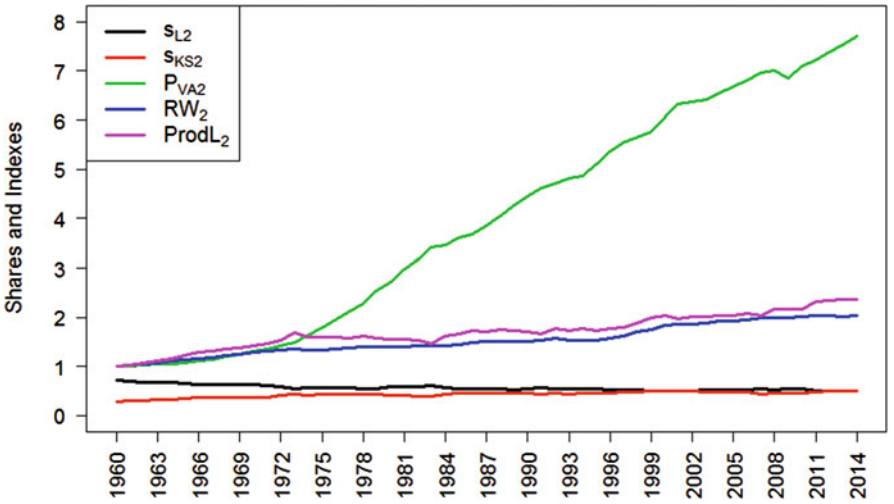


Fig. 2 Sector 2 labour and capital shares of value added, output price, real wage and labour productivity

4 Capital Stocks and Capital Output Ratios for Sectors 1 and 2

We constructed chained Fisher capital stock price and quantity indexes for Sector 1 using price and quantity information for each of the nine assets that are used as inputs, which are as follows: 1 = Equipment; 2 = Intellectual property products;

3 = Nonresidential structures; 4 = Residential structures; 5 = Residential land; 6 = Farm land; 7 = Commercial land; 8 = Beginning of year inventory stocks, and 9 = Beginning of the year real holdings of currency and deposits.

Denote the resulting period t price and quantity indexes as P_{K1}^t and Q_{K1}^t for t = 1960,...,2015. Define the Sector 1 capital stock value at the beginning of year t as $V_{K1}^t \equiv P_{K1}^t Q_{K1}^t$. Now define the year t nominal and real capital output ratios as $V_{K/O,1}^t \equiv V_{K1}^t / V_{VA1}^t$ and $Q_{K/O,1}^t \equiv Q_{K1}^t / Q_{VA1}^t$. V_{K1}^t, Q_{K1}^t, P_{K1}^t, $V_{K/O,1}^t$ and $Q_{K/O,1}^t$ are listed in Table 3.

It can be seen that the Sector 1 aggregate capital stock price P_{K1}^t increased 7.36 fold over the sample period. The average geometric growth rates for the price and quantity of the Sector 1 capital stock were 3.70% per year and 2.74% per year respectively. The real capital output ratio, $Q_{K/O,1}^t$, declined more or less steadily from 2.47 in 1960 to 1.59 in 2014. The nominal capital output ratio, $V_{K/O,1}^t$, did not decline nearly as much due to increasing land prices.[22] The nominal capital output ratio started at 2.47 and ended up at 2.52 with many fluctuations in between ($V_{K/O,1}^t$ had a low of 2.00 in 1966 and a high of 2.80 in 2009).

We similarly constructed chained Fisher capital stock price and quantity indexes for Sector 2 using the price and quantity information for each of the fourteen assets that are used as inputs,[23] which are as follows: 1 = Equipment held by sole proprietors; 2 = Equipment held by partners; 3 = Equipment held by cooperatives; 4 = Intellectual property products held by sole proprietors; 5 = Intellectual property products held by partners; 6 = Nonresidential structures held by sole proprietors; 7 = Nonresidential structures held by partners; 8 = Nonresidential structures held by cooperatives; 9 = Residential structures held by the noncorporate nonfinancial sector; 10 = Residential land held by the noncorporate nonfinancial sector; 11 = Farm land held by the noncorporate nonfinancial sector; 12 = Commercial land held by noncorporate nonfinancial sector; 13 = Beginning of the year inventories held by the noncorporate nonfinancial sector, and 14 = Beginning of the year real holdings of currency and deposits by noncorporate nonfinancial sector.

Denote the resulting beginning of period t price and quantity indexes as P_{K2}^t and Q_{K2}^t for t = 1960,...,2015. Define the Sector 2 capital stock value at the beginning of year t as $V_{K2}^t \equiv P_{K2}^t Q_{K2}^t$. Now define the year t nominal and real capital output ratios for Sector 2 as $V_{K/O,2}^t \equiv V_{K2}^t / V_{VA2}^t$ and $Q_{K/O,2}^t \equiv Q_{K2}^t / Q_{VA2}^t$. V_{K2}^t, Q_{K2}^t, P_{K2}^t, $V_{K/O,2}^t$ and $Q_{K/O,2}^t$ are listed in Table 3.

[22] We constructed chained Fisher land price and quantity indexes for Sector 1 and then compared the value of land to value added and the quantity of land to the quantity of output. The nominal land to output ratio went from 36.7% in 1960 to a peak of 51.2% in 2006, declined to 22.0% in 2012 and finished up in 2014 at 30.4%. The corresponding real land to output ratio declined steadily from 36.7% in 1960 to 9.8% in 2014. The inclusion or exclusion of land from the productive asset base does make a significant difference to capital output ratios.

[23] The BEA Fixed Asset Tables are organized somewhat differently for the Nonfinancial Noncorporate Sector as compared to Sector 1, with a decomposition of Sector 2 into subsectors. This led us to organize the capital stock data for Sector 2 into fourteen rather than nine components.

Table 3 Capital stock values, prices and quantities and nominal and real capital output ratios for Sectors 1 and 2

Year	$V_{K1}{}^t$	$Q_{K1}{}^t$	$P_{K1}{}^t$	$V_{K/O,1}{}^t$	$Q_{K/O,1}{}^t$	$V_{K2}{}^t$	$Q_{K2}{}^t$	$P_{K2}{}^t$	$V_{K/O,2}{}^t$	$Q_{K/O,2}{}^t$
1960	633.4	633.4	1.0000	2.4752	2.4752	368.9	368.9	1.0000	3.4345	3.4345
1961	641.2	650.0	0.9865	2.4399	2.4807	373.3	371.4	1.0049	3.3872	3.4248
1962	662.5	668.8	0.9906	2.3092	2.3533	385.6	376.4	1.0244	3.3762	3.3969
1963	689.9	691.8	0.9972	2.2539	2.2930	402.2	382.0	1.0530	3.4348	3.4002
1964	712.8	713.3	0.9993	2.1561	2.2090	411.5	384.3	1.0707	3.3455	3.3116
1965	750.5	739.3	1.0151	2.0579	2.1122	422.5	389.6	1.0844	3.2499	3.2397
1966	804.3	773.4	1.0400	1.9953	2.0572	445.9	399.5	1.1162	3.2197	3.2139
1967	874.7	816.8	1.0709	2.0634	2.1124	472.1	407.0	1.1601	3.3222	3.2916
1968	944.1	858.8	1.0994	2.0286	2.0855	500.5	412.4	1.2136	3.3409	3.3074
1969	1028.6	895.4	1.1487	2.0393	2.0909	548.1	422.7	1.2967	3.4733	3.3591
1970	1131.9	931.5	1.2152	2.1826	2.1942	614.3	436.1	1.4086	3.7596	3.4784
1971	1233.1	961.7	1.2823	2.2079	2.1794	675.1	445.3	1.5161	3.8998	3.5101
1972	1342.0	989.3	1.3565	2.1569	2.0841	746.7	459.5	1.6249	3.9052	3.4373
1973	1462.6	1020.1	1.4338	2.0933	2.0259	838.5	476.2	1.7609	3.7515	3.1803
1974	1663.0	1056.7	1.5738	2.2006	2.1309	1005.7	495.0	2.0320	4.2761	3.4374
1975	2016.4	1109.2	1.8180	2.4647	2.2689	1183.2	500.6	2.3634	4.6952	3.5795
1976	2216.9	1130.3	1.9613	2.3886	2.1373	1342.6	506.9	2.6486	4.8716	3.6025
1977	2449.5	1166.5	2.0999	2.3264	2.0544	1497.6	509.0	2.9422	4.9803	3.5939
1978	2725.9	1206.9	2.2587	2.2689	1.9958	1705.4	515.1	3.3106	5.0057	3.4396
1979	3098.1	1249.0	2.4806	2.3093	2.0013	2007.9	527.2	3.8085	5.2798	3.5257
1980	3569.1	1296.1	2.7537	2.4569	2.0974	2296.6	538.1	4.2683	5.7445	3.6514
1981	4058.3	1331.1	3.0488	2.4720	2.0700	2488.8	544.4	4.5719	5.7149	3.7201
1982	4528.4	1377.0	3.2886	2.6616	2.1892	2647.9	554.8	4.7728	5.8260	3.8671
1983	4779.9	1405.0	3.4020	2.6299	2.1295	2736.8	565.4	4.8408	5.6934	4.0070
1984	4966.4	1436.6	3.4570	2.4339	1.9993	2866.2	570.4	5.0250	5.1523	3.5380
1985	5280.1	1493.8	3.5346	2.4300	1.9868	2962.8	582.3	5.0879	4.9347	3.5092
1986	5540.5	1546.5	3.5826	2.4508	2.0046	3134.5	597.4	5.2473	4.9262	3.4568
1987	5742.8	1584.8	3.6237	2.3675	1.9508	3314.2	603.7	5.4899	4.9651	3.4934
1988	6041.1	1614.4	3.7420	2.2877	1.8722	3522.6	609.2	5.7828	4.8448	3.3930
1989	6434.3	1643.3	3.9156	2.3207	1.8701	3777.3	616.6	6.1259	4.8796	3.3988
1990	6749.5	1674.7	4.0303	2.3293	1.8787	4005.2	622.3	6.4360	4.9600	3.4255
1991	7057.4	1707.1	4.1343	2.3955	1.9250	4108.2	627.4	6.5480	5.0407	3.5639
1992	7229.2	1732.9	4.1716	2.3513	1.8963	4206.9	626.7	6.7131	4.8367	3.3879
1993	7421.0	1754.3	4.2302	2.3075	1.8735	4256.2	625.6	6.8033	4.7108	3.3327
1994	7775.5	1793.9	4.3345	2.2435	1.8047	4388.0	626.2	7.0077	4.6132	3.2049
1995	8173.3	1842.8	4.4353	2.2194	1.7686	4487.7	629.3	7.1314	4.5239	3.2346
1996	8648.2	1895.7	4.5619	2.2037	1.7187	4688.9	635.2	7.3817	4.3842	3.1841
1997	9059.0	1956.6	4.6299	2.1469	1.6629	4897.0	647.3	7.5658	4.3073	3.1541
1998	9577.5	2032.2	4.7129	2.1422	1.6343	5191.2	658.6	7.8816	4.2225	3.0334
1999	10131.9	2110.4	4.8009	2.1352	1.6096	5607.1	669.6	8.3735	4.2727	2.9424
2000	10861.3	2200.2	4.9366	2.1452	1.5909	6082.2	682.9	8.9060	4.2811	2.9169
2001	11758.9	2291.5	5.1316	2.3395	1.6929	6733.2	701.0	9.6050	4.1129	2.7133

(continued)

Table 3 (continued)

Year	V_{K1}^t	Q_{K1}^t	P_{K1}^t	$V_{K/O,1}^t$	$Q_{K/O,1}^t$	V_{K2}^t	Q_{K2}^t	P_{K2}^t	$V_{K/O,2}^t$	$Q_{K/O,2}^t$
2002	12238.7	2324.1	5.2661	2.4159	1.7024	7256.3	709.3	10.230	4.2509	2.6448
2003	12832.2	2348.3	5.4645	2.4542	1.6836	7847.9	710.4	11.048	4.3588	2.5323
2004	13575.4	2393.6	5.6715	2.4342	1.6422	8495.9	715.8	11.869	4.3486	2.3986
2005	14835.3	2432.0	6.1000	2.4896	1.6153	9558.0	726.9	13.149	4.5763	2.3262
2006	16327.0	2484.1	6.5727	2.5599	1.5887	10742.6	737.0	14.577	4.6847	2.1870
2007	17248.2	2519.9	6.8447	2.6247	1.5956	11334.8	750.9	15.094	4.8104	2.2219
2008	17652.5	2538.6	6.9537	2.6649	1.6284	11138.3	764.9	14.561	4.5012	2.1637
2009	17527.8	2561.5	6.8429	2.8027	1.7687	10275.4	775.7	13.247	4.4272	2.2938
2010	17154.9	2583.4	6.6405	2.5970	1.6901	9937.9	772.2	12.869	4.1486	2.2876
2011	17460.2	2620.2	6.6637	2.5225	1.6723	9464.8	773.8	12.232	3.6503	2.1555
2012	18064.0	2648.8	6.8197	2.4673	1.6250	9539.9	779.4	12.240	3.4788	2.0953
2013	18919.5	2688.7	7.0367	2.4921	1.6010	10491.4	788.3	13.309	3.6944	2.0912
2014	19908.8	2754.4	7.2281	2.5215	1.5896	11284.8	797.0	14.160	3.8043	2.0717
2015	20661.5	2806.7	7.3615			11960.3	810.2	14.762		

Note: All values are in billions of dollars and the quantities are in billions of 1960 dollars

It can be seen that the Sector 2 aggregate capital stock price P_{K2}^t increased 14.76 fold over the sample period whereas the Sector 1 capital stock price increased only 7.36 fold. The average geometric growth rates for the price and quantity of the Sector 2 capital stock were 5.02% per year (3.70% per year for Sector 1) and 1.44% per year (2.74% per year for Sector 1) respectively. This large difference in growth rates between sectors is explained by the relatively very large land component in the Sector 2 capital stock.[24] The price of land tends to grow more rapidly and the quantity less rapidly than other assets. The real capital output ratio for Sector 2, $Q_{K/O,2}^t$, increased (erratically) from 3.43 in 1960 to 4.01 in 1983 and then declined to 2.07 in 2014. The corresponding nominal capital output ratio, $V_{K/O,2}^t$, did not decline nearly as much, due to increasing land prices. The nominal capital output ratio started at 3.43 and remained roughly constant until 1969 and then increased rapidly to hit a peak of 5.83 in 1982 and then fell to 3.48 in 2012 and increased a little to end up at 3.80 in 2014. It can be seen that the real and nominal capital output ratios are in general, much larger in Sector 2 than in Sector 1.

The nominal and real capital output ratios for Sectors 1 and 2 are plotted in Fig. 3, where the overall decline in the real capital output ratios from 1983 is visible. The much higher capital output ratios for Sector 2 over Sector 1 are also apparent.

Figure 4 plots real value added, labour input and beginning of the period capital stocks for Sectors 1 and 2, except that each series is divided by its starting 1960 value. Thus real value added in Sectors 1 and 2 grew 6.77 fold and 3.58 fold

[24]The average share of residential land in Sector 2 value of the capital stock is 28.2%, farm land is 16.4% and commercial (nonresidential and nonfarm) land is 7.1%. Thus the overall average land share in the total value of Sector 2 assets is 51.6% and for reproducible assets is 48.4%. The average land share of asset value in Sector 1 is only 14.4% and the corresponding reproducible asset share is 85.6%.

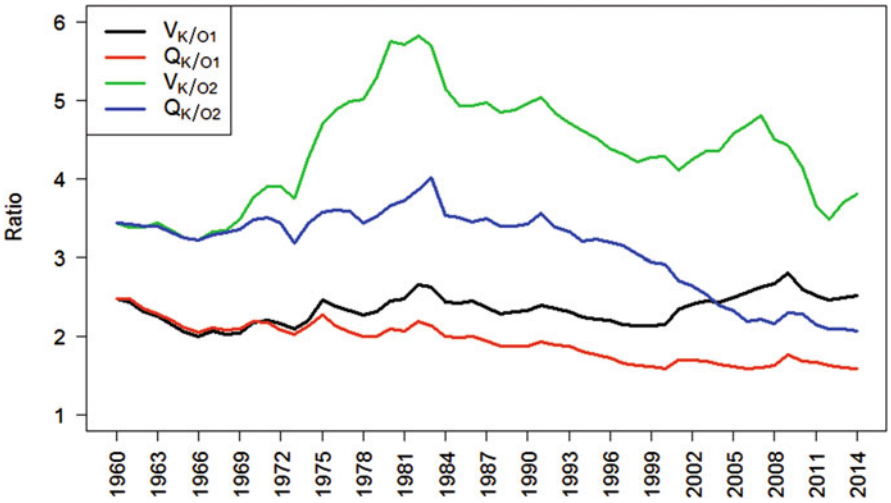


Fig. 3 Nominal and real capital output ratios for Sectors 1 and 2

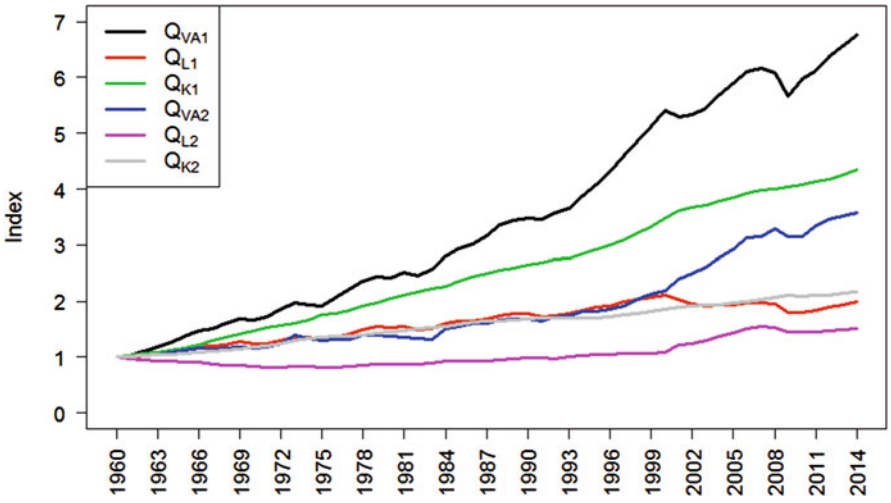


Fig. 4 Normalized and real value added, labour input and capital stocks for Sectors 1 and 2

respectively, labour input grew 1.98 fold in Sector 1 and 1.51 fold in Sector 2 and capital stocks grew 4.35 fold in Sector 1 and only 2.16 fold for Sector 2. Note that labour input in Sector 2 did not recover to its starting value in 1960 until 1993 after which it grew fairly rapidly until 2007 when it levelled off.

In the following section, we turn our attention to deriving the alternative balancing rates of return on assets, and the resulting user costs, for our two sectors that were discussed in Sect. 2.

5 Balancing Rates of Return and Alternative User Costs for Sectors 1 and 2

Denote the beginning of the year t asset prices for Sector 1 by $P_{K1,n}{}^t$ for n = 1,...,9. The year t inflation rate for asset n, $i_{1,n}{}^t$, is defined as follows:

$$i_{1,n}{}^t \equiv \left(P_{K1,n}{}^{t+1}/P_{K1,n}{}^t\right) - 1; \qquad n = 1,\ldots,9; t = 1960,\ldots,2014. \tag{16}$$

Denote the depreciation rate for asset n in year t used in Sector 1 by $\delta_{1,n}{}^t$. Define the depreciation rates for assets n = 5,...,9 to be 0 for all years t.[25]

Recall Eq. (9) in Sect. 2 which defined the ex post rate of return on assets for year t, r^t. For Sector 1, we will use the following counterpart to (9) to define the year t *ex post rate of return on assets for Sector 1*, $r_1{}^t$:

$$V_{VA1}{}^t - V_{L1}{}^t - \sum_{n=1}^{9}\left[1 + r_1{}^t - \left(1 + i_{1,n}{}^t\right)\left(1 - \delta_{1,n}{}^t\right)\right]P_{K1,n}{}^t Q_{K1,n}{}^t = 0;$$
$$t = 1960,\ldots,2014, \tag{17}$$

where Sector 1 value added and the value of labour input in year t, $V_{VA1}{}^t$ and $V_{L1}{}^t$, are listed in Table 1. The Sector 1 ex post rates of return on assets (the $r_1{}^t$ which solve (17) for year t data) are plotted in Fig. 5.

Recall that the personal consumption deflator for the beginning of year t was defined in Sect. 3 as $P_C{}^t$ for t = 1960,...,2014. Define the corresponding year t *consumption inflation rate*, $i_C{}^t$, by (18) and the corresponding year t *ex post real rate of return on assets* for Sector 1, $R_1{}^t$, by (19):

$$i_C{}^t \equiv \left(P_C{}^{t+1}/P_C{}^t\right) - 1; \qquad t = 1960\ldots,2014; \tag{18}$$

$$R_1{}^t \equiv \left[\left(1 + r_1{}^t\right)/\left(1 + i_C{}^t\right)\right] - 1; \qquad t = 1960\ldots,2014. \tag{19}$$

The personal consumption deflator inflation rates $i_C{}^t$ and the Sector 1 ex post real rates of return $R_1{}^t$ are also plotted in Fig. 5.

We also calculated a balancing rate of return for Sector 1 for each year t, $r_1{}^{t*}$, using a modification of Eq. (11) in Sect. 3. In order to calculate this alternative rate of return on assets, we need to form *expected or predicted asset inflation rates*, $i_{1,n}{}^{t*}$, for each asset n. For the first six years in our sample, we used the actual geometric

[25]The nonzero depreciation rates for assets n = 1,2,3,4 used in Sector 1 are listed in Table A10 in the Appendix of Diewert and Fox (2016).

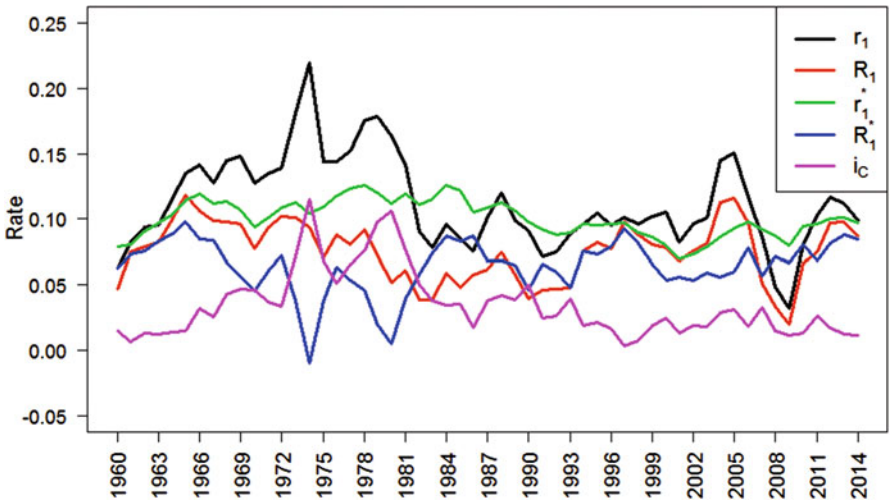


Fig. 5 Sector 1 nominal and real rates of return, predicted nominal and real rates of return, and personal consumption deflator inflation rate

average growth rate of the asset prices, starting at the beginning of 1960 and ending at the beginning of 1965. Thus we defined $i_{1,n}{}^{t^*}$ as follows for the first six years in our sample:

$$i_{1,n}{}^{t^*} \equiv \left(P_{K1,n}{}^{1965}/P_{K1,n}{}^{1960}\right)^{1/5} - 1; \qquad n = 1, \ldots, 9; t = 1960, \ldots, 1965. \tag{20}$$

For the years 1966–1985 we defined the $i_{1,n}{}^{t*}$ as geometric average growth rates of the asset price from the beginning of 1960 to the beginning of year t as follows for n = 1,...,9:

$$\begin{aligned} i_{1,n}{}^{1960*} &\equiv \left(P_{K1,n}{}^{1966}/P_{K1,n}{}^{1960}\right)^{1/6} - 1, \\ i_{1,n}{}^{1967*} &\equiv \left(P_{K1,n}{}^{1967}/P_{K1,n}{}^{1960}\right)^{1/7} - 1, \\ &\ldots \\ i_{1,n}{}^{1985*} &\equiv \left(P_{K1,n}{}^{1985}/P_{K1,n}{}^{1960}\right)^{1/25} - 1. \end{aligned} \tag{21}$$

For t greater than 1985, we simply used the geometric average growth rate of the asset price over the 25 years prior to year t; i.e., define $i_{1,n}{}^{t*}$ for $t \geq 1985$ as follows[26]:

[26]It may be that the length of our moving average process is too long or that better methods for predicting asset prices one year hence could be devised. However, our goal is to obtain user costs that could approximate one year rental prices for assets used in production (when they exist). Since

$$i_{1,n}{}^{t*} \equiv \left(P_{K1,n}{}^{t}/P_{K1,n}{}^{t-25}\right)^{1/25} - 1; \qquad n = 1,\ldots,9; t = 1985,\ldots,2014. \quad (22)$$

Recall Eq. (11) in Sect. 2 which decomposed value added into labour and capital service components using predicted asset inflation rates, which we now denote by $i_{1,n}{}^{t*}$, and a predicted or expected balancing nominal rate of return on assets for year t, which we now denote by $r_1{}^{t*}$. For Sector 1, we will use the following counterpart to (11) to define the year t *predicted balancing rate of return on assets for Sector 1*, $r_1{}^{t*}$:

$$V_{VA1}{}^{t} - V_{L1}{}^{t} - \sum_{n=1}^{9}\left[1 + r_1{}^{t*} - \left(1 + i_{1,n}{}^{t*}\right)\left(1 - \delta_{1,n}{}^{t}\right)\right]P_{K1,n}{}^{t}Q_{K1,n}{}^{t} = 0;$$

$$t = 1960,\ldots,2014 \quad (23)$$

where Sector 1 value added and the value of labour input in year t. The Sector 1 predicted rates of return on assets (the $r_1{}^{t*}$ which solve (23) for year t data) are plotted in Fig. 5.[27] The corresponding year t *predicted real rate of return on assets* for Sector 1, $R_1{}^{t*}$, is defined by (24) and also plotted in Fig. 5:

$$R_1{}^{t*} \equiv \left[\left(1 + r_1{}^{t*}\right)/\left(1 + i_C{}^{t}\right)\right] - 1; \qquad t = 1960\ldots,2014. \quad (24)$$

The mean nominal rate of return $r_1{}^{t}$ over the sample period in Sector 1 was 11.25% (minimum rate was 3.21% in 2009 and the maximum was 21.97% in 1974) while the mean real ex post rate of return on assets $R_1{}^{t}$ was 7.57% (minimum was 1.99% in 2009; maximum was 11.83% in 1965). These ex post real rates have been above average for the last three years at 9.73%, 9.82% and 8.68%. The mean nominal predicted rate of return $r_1{}^{t*}$ over the sample period in Sector 1 was 10.04% (minimum rate was 6.96% in 2001 and the maximum was 12.56% in 1978) while the mean expected real rate of return on assets $R_1{}^{t*}$ was 6.44% (minimum was −0.94% in 1974; maximum was 9.77% in 1965).[28]

The most important series is $R_1{}^{t}$, the before income tax realized real rate of return on assets used in the Corporate Nonfinancial Sector.[29] This real rate has remained above 5% except for the 10 years 1960, 1982–83, 1985, 1990–93 and 2008–09, and has remained below 11% except for the 3 years 1965 and 2004–05. There is no indication of a real rate of return slowdown that shows up in our data. However, the 2008 financial crisis certainly drove down ex post realized rates of return temporarily in 2008 and 2009.

observed rental prices are relatively smooth, our suggested method for generating predicted asset prices does lead to relatively smooth user costs as will be seen later.

[27] Tabulated data for the series in this and following figures are available in Diewert and Fox (2016).

[28] Note that our expected real rate of return on Sector 1 assets has been fairly stable over the period 1982–2014. $R_1{}^{t*}$ ranged between 4.62% (1990) and 9.33% (1997) over this period.

[29] The average corporate income tax paid by the nonfinancial corporate sector on assets during our sample period as a percentage of the asset base is 1.98% per year; see the series $V_{TI1}{}^{t}$ in Appendix Table A3 of Diewert and Fox (2016).

We turn our attention to Sector 2. Denote the beginning of the year t asset prices for Sector 2 by $P_{K2,n}{}^t$ for n = 1,...,14. The year t inflation rate for asset n in Sector 2, $i_{2,n}{}^t$, is defined as follows:

$$i_{2,n}{}^t \equiv \left(P_{K2,n}{}^{t+1}/P_{K2,n}{}^t\right) - 1; \qquad n = 1,\ldots,14; t = 1960,\ldots,2014. \tag{25}$$

Denote the depreciation rate for asset n in year t used in Sector 2 by $\delta_{2,n}{}^t$. Define the depreciation rates for assets n = 10,...,14 to be 0 for all years t.[30] Again recall Eq. (9) in Sect. 2 which defined the ex post rate of return on assets for year t, r^t. For Sector 2, we will use the following counterpart to Eq. (9) to define the year t *ex post rate of return on assets for Sector 2*, $r_2{}^t$:

$$V_{VA2}{}^t - V_{L2}{}^t - \sum_{n=1}^{14}\left[1 + r_2{}^t - \left(1 + i_{2,n}{}^t\right)\left(1 - \delta_{2,n}{}^t\right)\right]P_{K2,n}{}^t Q_{K2,n}{}^t = 0;$$
$$t = 1960,\ldots,2014 \tag{26}$$

where Sector 2 value added and the value of labour input in year t, $V_{VA2}{}^t$ and $V_{L2}{}^t$, are listed in Table 2. The Sector 2 ex post rates of return on assets (the $r_2{}^t$ which solve (26) for year t data) are plotted in Fig. 6. The year t *ex post real rate of return on assets* for Sector 2, $R_2{}^t$, is defined by (27):

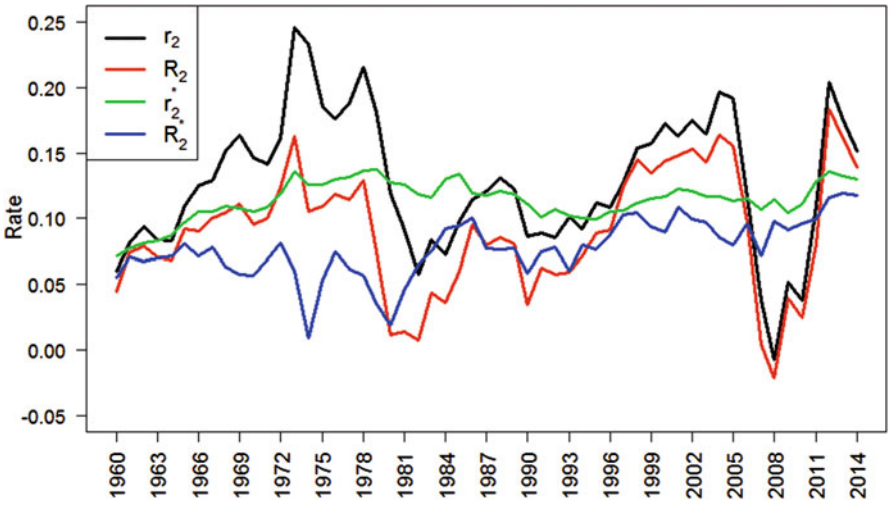


Fig. 6 Sector 2 nominal and real rates of return, predicted nominal and real rates of return

[30]The nonzero depreciation rates for assets n = 1,...,9 used in Sector 2 are listed in Table A11 in the Appendix of Diewert and Fox (2016).

$$R_2{}^t \equiv \left[(1 + r_2{}^t) / \left(1 + i_C{}^t\right)\right] - 1; \qquad t = 1960 \ldots, 2014. \tag{27}$$

We also calculated a balancing rate of return for Sector 2 for each year t, $r_2{}^{t*}$, using a modification of Eq. (11) in Sect. 3. In order to calculate this alternative rate of return on assets, we need to form *expected or predicted asset inflation rates*, $i_{2,n}{}^{t*}$, for each asset n. We formed Sector 2 predicted asset inflation rates using exactly the same method that we used to form Sector 1 predicted inflation rates.

Recall Eq. (11) in Sect. 2 which decomposed value added into labour and capital service components using predicted asset inflation rates, which we now denote by $i_{2,n}{}^{t*}$, and a predicted or expected balancing nominal rate of return on assets for year t, which we now denote by $r_2{}^{t*}$. For Sector 2, we will use the following counterpart to (11) to define the year t *predicted balancing rate of return on assets for Sector 2*, $r_2{}^{t*}$:

$$V_{VA2}{}^t - V_{L2}{}^t - \sum_{n=1}^{14} \left[1 + r_2{}^{t*} - \left(1 + i_{2,n}{}^{t*}\right)\left(1 - \delta_{2,n}{}^t\right)\right] P_{K2,n}{}^t Q_{K2,n}{}^t = 0;$$
$$t = 1960, \ldots, 2014 \tag{28}$$

where Sector 2 value added and the value of labour input in year t, $V_{VA2}{}^t$ and $V_{L2}{}^t$, are listed in Table 2. The Sector 2 predicted rates of return on assets (the $r_2{}^{t*}$ which solve (28) for year t data) are plotted in Fig. 6, along with the corresponding year t *predicted real rate of return on assets* for Sector 2, $R_2{}^{t*}$, as defined by (29):

$$R_2{}^{t*} \equiv \left[\left(1 + r_2{}^{t*}\right) / \left(1 + i_C{}^t\right)\right] - 1; \qquad t = 1960 \ldots, 2014. \tag{29}$$

The mean nominal ex post rate of return $r_2{}^t$ over the sample period in Sector 2 was 12.76% (minimum rate was −0.70% in 2008 and the maximum was 24.60% in 1973) while the mean real ex post rate of return on assets $R_2{}^t$ was 9.03% (minimum was −2.14% in 2008; maximum was 18.29% in 2012). Note that the average real rate of return in Sector 2 was a very high 9.03% per year which is considerably above the average real rate of return on assets used in Sector 1, which was 7.57% per year. This result was somewhat surprising. The Sector 2 ex post real rates have been above average for the last 3 years at 18.29%, 16.13% and 13.86%. These are very high real rates of return. The corresponding Sector 1 ex post real rates were only 9.73%, 9.82% and 8.68%.[31] The mean *nominal predicted* rate of return $r_n{}^{t*}$ over the sample period in Sector 2 was 11.35% (minimum rate was 7.12% in 1960 and the maximum was 13.72% in 1979) while the mean expected *real predicted* rate of return on assets $R_1{}^{t*}$ was 7.73% (minimum was 0.92% in 1974; maximum was 11.91% in 2013).

[31]The reason why nominal and real ex post rates of return on assets are much higher in Sector 2 compared to Sector 1 can be explained by the fact that production in Sector 2 is highly land intensive and land inflation rates are much higher than inflation rates for other assets.

The most important series is R_2^t, the before income tax realized real rate of return on assets used in the Noncorporate Nonfinancial Sector.[32] This series has fluctuated considerably during the sample period, driven by large fluctuations in the price of land. There does not appear to be a long run decline in the real rate of return on assets in Sector 2. The predicted nominal rate of return series r_2^{t*} is much smoother than the corresponding realized return series r_2^t and so the use of the r_2^{t*} series in our user costs will lead to much smoother user costs for this sector.

We turn our attention to the calculation of user costs for Sector 1. Recall Eqs. (16) and (17). The *year t Jorgensonian user cost for asset n used in Sector 1*, $u_{1,n}^t$, is defined as follows:

$$u_{1,n}^t \equiv \left[1 + r_1^t - \left(1 + i_{1,n}^t\right)\left(1 - \delta_{1,n}^t\right)\right] P_{K1,n}^t; \quad n = 1, \ldots, 9; t = 1960, \ldots, 2014 \tag{30}$$

where the $i_{1,n}^t$ are the ex post asset inflation rates defined by (16) and the r_1^t are the Sector 1 balancing nominal rates of return defined by eqs. (17). These Jorgensonian user costs are plotted in Fig. 7. It can be seen that there are numerous negative Jorgensonian user costs for assets 5–8 (residential land, farm land, commercial land and inventory stocks). It can also be seen that these user costs are in general quite volatile. Thus while Jorgensonian user costs are the "right" user costs to use when

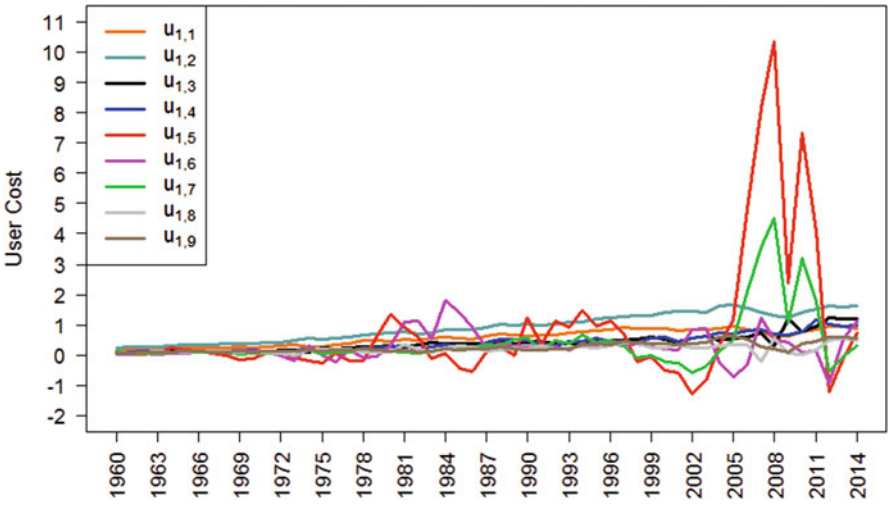


Fig. 7 Jorgensonian user costs for Sector 1

[32]The average business income tax paid by the nonfinancial noncorporate sector on assets during our sample period as a percentage of the asset base is only 0.15% per year; see the series V_{T12}^t in Appendix Table A3 of Diewert and Fox (2016). This income tax rate for Sector 2 seems to be too low to be true!

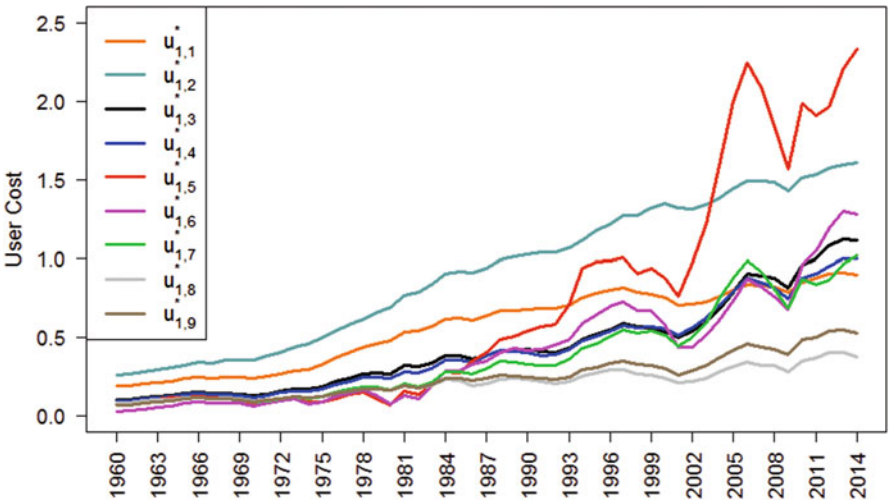


Fig. 8 Predicted user costs for Sector 1

computing ex post rates of return on assets, they are not good approximations to rental prices for these assets.[33]

Recall Eqs. (20, 21, 22 and 23). The *year t predicted user cost for asset n used in Sector 1*, $u_{1,n}{}^{t^*}$, is defined as follows:

$$u_{1,n}{}^{t^*} \equiv \left[1+r_1{}^{t^*}-\left(1+i_{1,n}{}^{t^*}\right)\left(1-\delta_{1,n}{}^{t}\right)\right]P_{K1,n}{}^{t};\ n=1,\ldots,9;t=1960,\ldots,2014 \tag{31}$$

where the $i_{1,n}{}^{t^*}$ are the predicted asset inflation rates defined by (20, 21 and 22) and the $r_1{}^{t^*}$ are the predicted Sector 1 balancing nominal rates of return defined by Eq. (23). These predicted user costs are plotted in Fig. 8.

The predicted user costs are much smoother than the Jorgensonian user costs and the negative user costs have been eliminated. Thus in what follows, we will sometimes refer to these predicted user costs as *smoothed user costs*. These user costs are suitable for production or cost function econometric studies. They are also more suitable for statistical agencies to use when computing capital services aggregates for publication. It can be seen that the user costs for residential, farm and commercial land ($u_{1,5}{}^{t^*}, u_{1,6}{}^{t^*}$ and $u_{1,7}{}^{t^*}$) have been quite volatile for the last 20 years in our sample period but the remaining user cost series are fairly smooth.

[33]Thus the use of Jorgensonian user costs is not recommended in econometric studies where cost functions are estimated or where production functions are estimated using inverse factor demand equations as additional estimating equations.

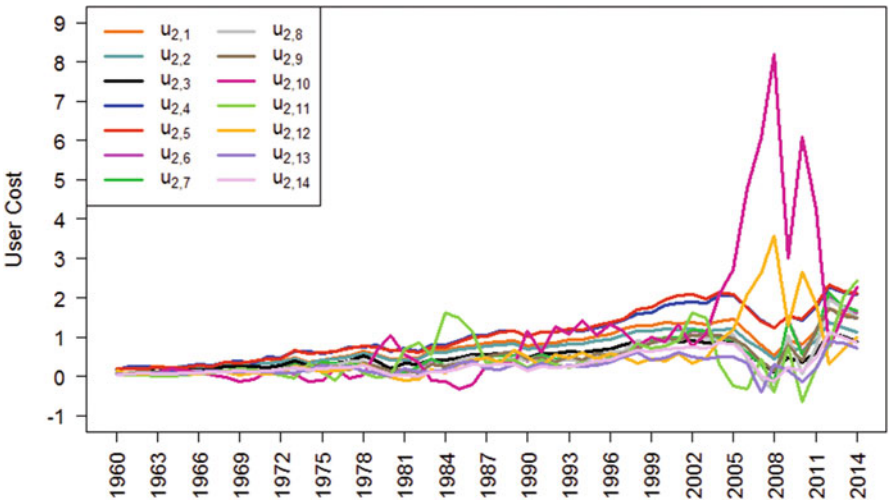


Fig. 9 Jorgensonian user costs for Sector 2

We turn our attention to the calculation of user costs for Sector 2. Recall Eq. (26). The *year t Jorgensonian user cost for asset n used in Sector 2*, $u_{2,n}{}^t$, is defined as follows:

$$u_{2,n}{}^t \equiv \left[1 + r_2{}^t - \left(1 + i_{2,n}{}^t\right)\left(1 - \delta_{2,n}{}^t\right)\right] P_{K2,n}{}^t; \quad n = 1, \ldots, 14; t = 1960, \ldots, 2014 \tag{32}$$

where the $i_{2,n}{}^t$ are the ex post asset inflation rates defined by (25) and the $r_2{}^t$ are the Sector 2 balancing nominal rates of return defined by Eq. (26). These Jorgensonian user costs are plotted in Fig. 9. Again, these user costs are volatile and there are numerous negative user costs in assets, 6–7 (nonresidential structures held by proprietors, partners and cooperatives) and 10–14 (residential land, farm land, commercial land, inventory stocks and monetary stocks). It can be seen at a glance that these user costs are not suitable approximations to asset rental prices.

Recall Eq. (28). The *year t predicted user cost for asset n used in Sector 2*, $u_{2,n}{}^{t*}$, is defined as follows:

$$u_{2,n}{}^{t*} \equiv \left[1 + r_2{}^{t*} - \left(1 + i_{2,n}{}^{t*}\right)\left(1 - \delta_{2,n}{}^t\right)\right] P_{K2,n}{}^t; \; n = 1, \ldots, 14; t = 1960, \ldots, 2014 \tag{33}$$

where the $i_{2,n}{}^{t*}$ are the predicted asset inflation rates for Sector 2 defined by counterparts to definitions (20, 21 and 22) and the $r_2{}^{t*}$ are the predicted Sector 2 balancing nominal rates of return defined by Eq. (28). These predicted user costs are plotted in Fig. 10. It can be seen that these predicted user costs are all positive, and that all of the series have fairly smooth trends, with the exception of assets 10, 11 and 12 (residential land, farm land and commercial land).

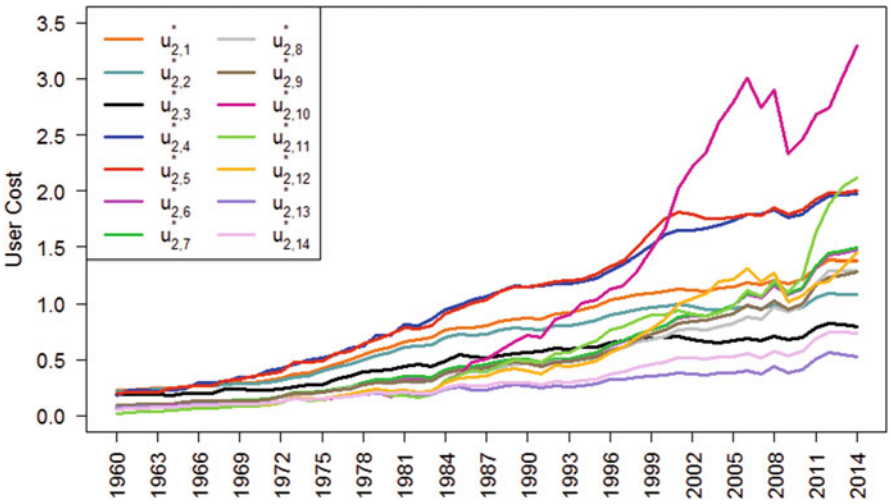


Fig. 10 Predicted user costs for Sector 2

We conclude that our rather simple method for forming predicted asset inflation rates does lead to relatively smooth (and reasonable) user costs that could be published by statistical agencies for general use by economic analysts as well as for the construction of capital services aggregates. In the following section, we will compute capital services aggregates (and the resulting measures of Total Factor Productivity) using both Jorgensonian and predicted user costs to determine if the alternative user costs affect aggregate capital services growth for our two sectors.

6 Jorgensonian and Predicted Measures of Capital Services and Total Factor Productivity Growth

We use the Törnqvist formula to aggregate capital services and to aggregate all inputs, including labour services.[34] Our methodology for measuring Total Factor Productivity growth follows the methodology proposed by Diewert and Morrison

[34]This formula was attributed to Törnqvist (1936) by Jorgenson and Griliches (1972; 83) as a discrete time approximation to the continuous time Divisia indexes that Jorgenson and Griliches (1967, 1972) advocated for aggregating inputs and outputs in productivity studies. The formula does not explicitly appear in Törnqvist (1936) but it is explicit in a follow up paper co-authored by Törnqvist; see Törnqvist and Törnqvist (1937). The formula was derived in an instructive manner by Theil (1967; 136–137) and so it is also known as the Törnqvist-Theil formula. Jorgenson and Nishimizu (1982) called the index the translog index. Diewert (1976; 118–129), Diewert and Morrison (1986) and Kohli (1990) related Törnqvist price and quantity indexes to various translog functional forms for cost, revenue and production functions.

(1986) and Kohli (1990). This methodology measures TFP growth over two periods as an implicit Törnqvist quantity index defined over gross outputs and intermediate inputs divided by a direct Törnqvist quantity index of primary inputs.[35] Since we have only one value added output in our BEA data base for each sector, our output index going from year t to year t + 1 is simply $Q_{VA1}{}^{t+1}/Q_{VA1}{}^{t}$ for Sector 1 and $Q_{VA2}{}^{t+1}/Q_{VA2}{}^{t}$ for Sector 2. However, we will use the Törnqvist quantity index to aggregate inputs.

Let $p^t \equiv [p_1{}^t, \ldots, p_N{}^t]$ and $q^t \equiv [q_1{}^t, \ldots, q_N{}^t]$ denote a generic price and quantity vector for year t. Then the logarithm of the *Törnqvist chain link quantity index* Q_T going from year t to t + 1 is defined as follows:

$$\ln Q_T\left(p^t, p^{t+1}, q^t, q^{t+1}\right) \equiv \sum_{n=1}^{N} (1/2)\left(s_n{}^t + s_n{}^{t+1}\right) \ln \left(q_n{}^{t+1}/q_n{}^t\right) \tag{34}$$

where the cost share of input n in year t is defined as $s_n{}^t \equiv p_n{}^t q_n{}^t/p^t \cdot q^t$ for n = 1,...,N. Note that this index can be used to aggregate quantities as long as they are all positive even though some prices may be negative.

The Törnqvist quantity index was used to aggregate the nine types of capital services used by Sector 1. Denote the aggregate *chained Törnqvist quantity index of Jorgensonian capital services* and of *predicted capital services* for Sector 1 for year t by $Q_{KJ1}{}^t$ and $Q_{KP1}{}^t$ respectively.[36] The Törnqvist quantity index was also used to aggregate the nine types of capital services and the one type of labour used by Sector 1. Denote the chained index for year t using Jorgensonian and predicted user costs by $Q_{XJ1}{}^t$ and $Q_{XP1}{}^t$ respectively.[37] Finally, the *year t levels of Jorgensonian and Predicted TFP* are defined as follows:

$$TFP_{J1}{}^t \equiv \left[Q_{VA1}{}^t/Q_{VA1}{}^{1960}\right]/Q_{XJ1}{}^t/Q_{XJ1}{}^{1960}\Big]; \qquad t = 1960, \ldots, 2014; \tag{35}$$

$$TFP_{P1}{}^t \equiv \left[Q_{VA1}{}^t/Q_{VA1}{}^{1960}\right]/Q_{XP1}{}^t/Q_{XP1}{}^{1960}\Big]; \qquad t = 1960, \ldots, 2014. \tag{36}$$

The quantity and TFP series are plotted in Fig. 11, along with the labour input series $Q_{L1}{}^t$ (normalized to equal 1 in 1960).

It can be seen that labour input into the Corporate Nonfinancial Sector grew fairly steadily to a 2.11 fold increase in 2000 but then growth levelled off and fell to

[35] See Diewert (2014b) for a detailed explanation of the methodology and an application to US data. The land data used in this earlier study was of lower quality than the land data used in the current study.

[36] These series are normalized to equal one in 1960.

[37] These series were also normalized to equal one in 1960. The price and value of labour input for Sector 1 in year t, $P_{L1}{}^t$ and $V_{L1}{}^t$, are listed in Table 1. Define the quantity of labour used in Sector 1 in year t as $Q_{L1}{}^t \equiv V_{L1}{}^t/P_{L1}{}^t$. Thus we added $P_{L1}{}^t$ and $Q_{L1}{}^t$ to our user costs and capital stock quantities to form the overall chained Törnqvist input quantity indexes.

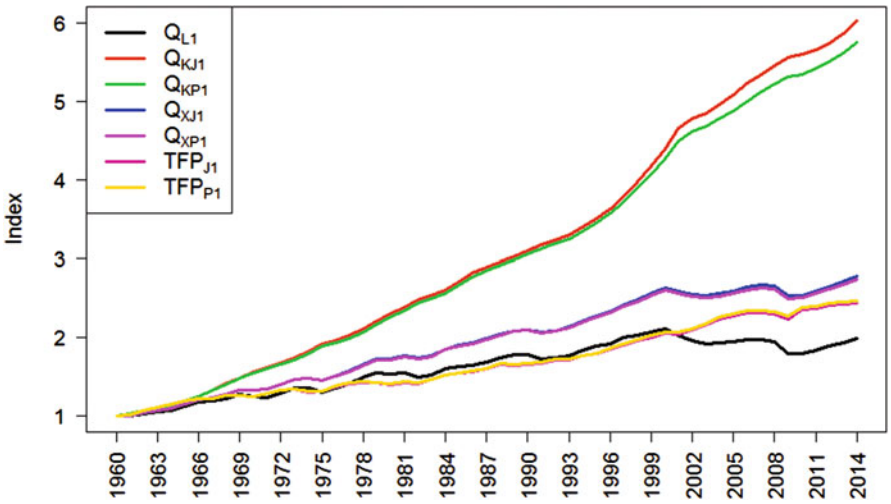


Fig. 11 Sector 1 indexes of labour quantity, and alternative capital services, aggregate input and TFP estimates

a 1.79 fold increase over 1960 in 2009 and 2010. Labour input has since increased to finish off at a 1.98 fold increase over 1960 in 2014. We note that the price of labour has increased steadily (even through the Great Recession period) to end up increasing 13.95 fold over the sample period. The geometric average rate of growth of Q_{L1}^t was 1.28% per year and the geometric average rate of growth of P_{L1}^t over the sample period was 5.00% per year.

The quantity of Jorgensonian capital services increased 6.02 fold over the sample period while the quantity of predicted capital services increased only 5.75 fold. The geometric average rates of growth for these two measures of capital services were 3.38% and 3.29% per year. This difference is surprisingly small considering how different the two sets of user costs were. The price index of Jorgensonian capital services increased 6.37 fold over the sample period while the price index of predicted capital services increased 6.68 fold. The geometric average rates of growth for these two measures of capital services prices were 3.49% and 3.58% per year. One reason why there is so little difference between the two measures of capital services is that land as a share of total capital services in Sector 1 is relatively small.[38]

Sector 1 Jorgensonian input Q_{XJ1}^t increased 2.78 fold over the sample period while the quantity of predicted capital services Q_{XP1}^t increased 2.74 fold.[39] The

[38] Using Jorgensonian user costs, we find that the sample average input cost shares of labour, land services and reproducible capital stock services in Sector 1 were 68.6%, 2.1% and 29.3%. The sample average cost shares of residential, farm and commercial land (assets 5, 6 and 7) were only 0.05%, 0.17% and 1.85%.

[39] Note that Q_{XJ1}^t and Q_{XP1}^t (and TFP_{J1}^t and TFP_{P1}^t) cannot be distinguished in Figure 11.

geometric average rates of growth for these two input measures were 1.91% and 1.89% per year. This is a very small difference in growth rates. Sector 1 real value added Q_{VA1}^t grew 6.77 fold over the sample period (geometric average rate of growth was 3.61% per year). Jorgensonian TFP in Sector 1, TFP_{J1}^t, grew 2.43 fold over the sample period while predicted TFP, TFP_{P1}^t, grew 2.47 fold. The geometric average rates of growth for these two measures of Total Factor Productivity were 1.66% and 1.69% per year, a surprisingly small difference.

Another surprise is the rather high overall rate of TFP growth that the Corporate Nonfinancial Sector has been able to achieve over the 55 years in our sample. To see if there has been a TFP slowdown over the past 15 years, we computed decade by decade geometric average rates of TFP growth.[40] Using Jorgensonian estimates for input growth, the resulting decade by decade averages were as follows: 2.57% (1960s), 1.22% (1970s), 1.51% (1980s), 1.99% (1990s), 1.09% (2000s) and 1.71% (2010s) per year.[41] There is little evidence of a productivity slowdown in Sector 1 using these sub-periods; the average TFP growth rate over the last 5 years in our sample is 1.71% per year, which is slightly higher than long run Jorgensonian average of 1.66% per year. However, if we consider the sub-period 2005–2014, the geometric average was 0.88%, which is substantially lower than the long run average.[42]

We turn our attention to developing alternative measures of capital services and productivity growth for Sector 2, the Noncorporate Nonfinancial Sector of the US private sector.

Again, the Törnqvist quantity index was used to aggregate the fourteen types of capital services used by Sector 2. Denote the aggregate *chained Törnqvist quantity index of Jorgensonian capital services* and of *predicted capital services* for Sector 2 for year t by Q_{KJ2}^t and Q_{KP2}^t respectively.[43] The Törnqvist quantity index was also used to aggregate the fourteen types of capital services and the one type of labour used by Sector 2. Denote the chained index for year t using Jorgensonian and predicted user costs by Q_{XJ2}^t and Q_{XP2}^t respectively.[44] Finally, the *year t levels of Jorgensonian and Predicted TFP* are defined as follows:

[40] The last "decade" covers only the years 2010–2014.

[41] Using predicted user costs, the corresponding decade by decade geometric average rates of TFP growth in Sector 1 were as follows: 2.59%, 1.26%, 1.49%, 2.02%, 1.16% and 1.72% per year.

[42] See Diewert and Fox (2017) on potential sources of the productivity slowdown.

[43] These series are normalized to equal one in 1960 when they are listed in Table 13. The input price and quantity series used in the index number formula for Q_{KJ2}^t and Q_{KP2}^t are the $u_{2,n}^t$ and $u_{2,n}^{t*}$ listed in Tables 9 and 11 respectively and the corresponding quantity series $Q_{K2,n}^t$ are described in Table 5.

[44] These series were also normalized to equal one in 1960. The price and value of labour input for Sector 1 in year t, P_{L2}^t and V_{L2}^t, are listed in Table 2. Define the quantity of labour used in Sector 2 in year t as $Q_{L2}^t \equiv V_{L2}^t/P_{L2}^t$. Thus we added P_{L2}^t and Q_{L2}^t to our user costs and capital stock quantities to form the overall chained Törnqvist input quantity indexes.

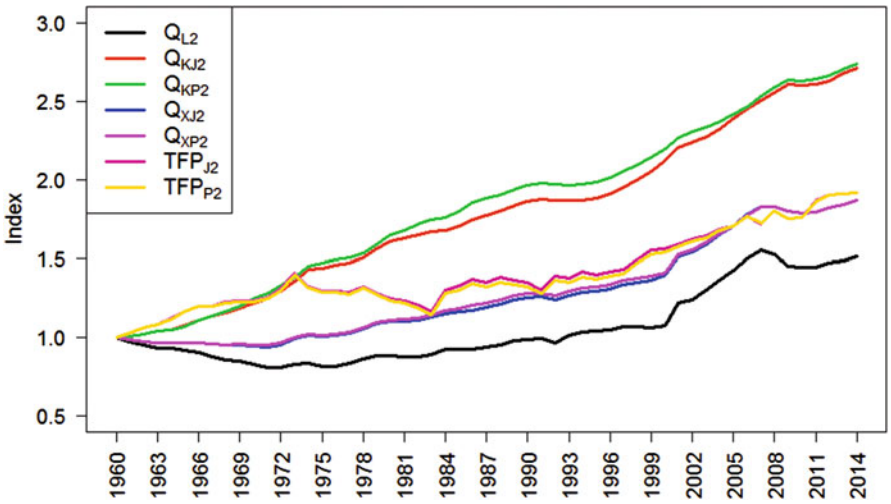


Fig. 12 Sector 2 indexes of labour quantity, and alternative capital services, aggregate input and TFP estimates

$$TFP_{J2}{}^{t} \equiv \left[Q_{VA2}{}^{t}/Q_{VA2}{}^{1960}\right]/\left[Q_{XJ2}{}^{t}/Q_{XJ2}{}^{1960}\right]; \qquad t = 1960, \ldots, 2014; \quad (37)$$

$$TFP_{P2}{}^{t} \equiv \left[Q_{VA2}{}^{t}/Q_{VA2}{}^{1960}\right]/\left[Q_{XP2}{}^{t}/Q_{XP2}{}^{1960}\right]; \qquad t = 1960, \ldots, 2014. \quad (38)$$

These quantity and TFP series, along with the labour input series $Q_{L2}{}^{t}$ (normalized to equal 1 in 1960), are plotted in Fig. 12. The rates of input, output and productivity growth in Sector 2 are quite different from the corresponding rates in Sector 1 as can be seen by comparing Figs. 11 and 12.

Labour input into the Noncorporate Nonfinancial Sector fell to 80.7% of its initial 1960 level in 1972 but then grew fairly steadily to a 1.07 fold increase in 2000 over its initial level. Then labour input growth grew rapidly to a 1.55 fold increase in 2007, fell to 1.44 in 2010 and then slowly increased to finish up with a 1.51 fold increase over its initial level. We note that the price of labour $P_{L2}{}^{t}$ has increased steadily to end up increasing 12.67 fold over the sample period. The geometric average rate of growth of $Q_{L2}{}^{t}$ was only 0.77% per year (compared to a 1.28% geometric rate of increase for $Q_{L1}{}^{t}$) and the geometric average rate of growth of $P_{L2}{}^{t}$ over the sample period was 4.81% per year, which is close to the rate of increase for $P_{L1}{}^{t}$ (5.00% per year).

The quantity of Jorgensonian capital services increased 2.71 fold in Sector 2 over the sample period (the Sector 1 increase was 6.02 fold) while the quantity of Sector 2 predicted capital services increased 2.74 fold. The geometric average rates of growth for these two measures of capital services were 1.86% and 1.88% per year (compared to 3.38% and 3.29% per year for Sector 1). Again, this difference

in average rates of capital services growth is surprisingly small considering how different the two sets of user costs were.[45] The price index for Jorgensonian capital services increased 17.94 fold over the sample period while the price index of predicted capital services increased 17.75 fold (only 6.37 fold and 6.68 fold increases for Sector 1 capital service prices). The geometric average rates of growth for these two measures of Sector 2 capital services prices were 5.49% and 5.47% per year (the corresponding rates for Sector 1 were 3.49% and 3.58% per year). Thus since land is a much more important input in Sector 2 compared to Sector 1, the overall rate of growth in the price of capital services in Sector 2 is much greater than in Sector 1.[46] Note that these rates of service price increase for Sector 2 are higher than the rate of increase in wages for Sector 2, which was only 4.81% per year.[47] Looking at Fig. 12, it can be seen that the level of predicted capital services, $Q_{KP2}{}^t$, bulged above the corresponding level of Jorgensonian capital services, $Q_{KJ2}{}^t$, over the middle of the sample period but the two series were quite close near the endpoints of our sample period.

Sector 2 Jorgensonian input $Q_{XJ2}{}^t$ and predicted input $Q_{XP2}{}^t$ increased 1.87 fold over the sample period (the Sector 1 counterparts were 2.78 fold and 2.74 fold increases).[48] The geometric average rates of growth for these two input measures were both 1.17% per year (1.91% and 1.89% for Sector 1). Sector 2 real value added $Q_{VA2}{}^t$ grew 3.58 fold (6.77 fold for Sector 1) over the sample period and the geometric average rate of growth was 2.39% per year (3.61% for Sector 1). Jorgensonian TFP and predicted TFP in Sector 2, $TFP_{J2}{}^t$ and $TFP_{J2}{}^t$, both grew 1.91 fold over the sample period (2.43 and 2.47 for Sector 1). The geometric average rates of growth for the two Sector 2 measures of Total Factor Productivity were both 1.21% per year (1.66% and 1.69% per year for Sector 1).

Another surprise is the rather high overall rate of TFP growth that the Noncorporate Nonfinancial Sector has been able to achieve over the 55 years in our sample. To see if there has been a TFP slowdown over the past 15 years, we computed decade

[45]However, the predicted asset price inflation rates are on average quite close to the average ex post asset price inflation rates. Thus on average, the two sets of user costs are similar, giving rise to similar trends in the two sets of capital service prices.

[46]Using Jorgensonian and predicted user costs, we find that the sample average input cost shares of labour and capital services were 56.7% and 43.3%. Using Jorgensonian user costs, the sample average cost shares of residential, farm and commercial land services (assets 10, 11 and 12) were 7.51%, 4.44% and 2.43%. Using predicted user costs, the sample average input cost shares for assets 10, 11 and 12 were 8.04%, 4.05% and 2.44%. These input cost shares for land are low compared to the share of land assets in total asset value: the average overall land share of total asset value was 51.6% while reproducible assets contributed 48.4% of total asset value. The average shares of the three types of land in total asset value were 28.2%, 16.4% and 7.1%. The user cost shares of capital services for land are lower than their corresponding asset value shares because the high land price inflation terms dramatically reduce land user costs relative to their asset prices.

[47]These trends in the prices and quantities of labour and capital input into Sector 2 indicate the presence of labour saving technical progress in this sector.

[48]Note that $Q_{XJ2}{}^t$ and $Q_{XP2}{}^t$ (and $TFP_{J2}{}^t$ and $TFP_{P2}{}^t$) can hardly be distinguished in Figure 12.

by decade geometric average rates of TFP growth.[49] Using Jorgensonian estimates for input growth, the resulting decade by decade $TFP_{J2}{}^t$ averages were as follows: 2.32% (1960s), 0.41% (1970s), 0.64% (1980s), 1.29% (1990s), 1.22% (2000s) and 1.81% (2010s) per year.[50] Thus there is little evidence of a productivity slowdown in Sector 2 using these sub-periods; the average Jorgensonian TFP growth rate for Sector 2 over the last 5 years in our sample is 1.81% per year, which is slightly higher than the corresponding Jorgensonian rate of 1.71% for Sector 1 over the past 5 years and higher than the long run Jorgensonian average of 1.66% per year for Sector 1. However, if we consider the sub-period 2005–2014, the geometric average was 1.27%, which is substantially lower than the long run average.

Finally, we note that for the period 2000–2009, Jorgensonian TFP growth averaged 1.22% per year while the corresponding predicted TFP growth averaged 1.37% per year. This is a substantial difference. Thus, although for the most part Jorgensonian TFP growth rates based on the use of ex post asset inflation rates are close to our preferred TFP growth rates based on the use of predicted asset inflation rates, it can be seen that it is not always the case that these rates are close.

In the following section, we look at what happens to the rate of return on assets and on Jorgensonian TFP growth rates when we drop assets from the asset boundary.

7 Rates of Return and TFP Growth in Sector 1 with Alternative Asset Bases

Many national and international productivity data bases do not include money, inventories or land in their asset base.[51] Thus it is of interest to see what happens to rates of return on assets and on TFP growth when these assets are dropped from the list of productive inputs.

Recall Eqs. (17) and (19) in Sect. 5 which defined the year t nominal and real rate of return on all nine assets used in Sector 1, $r_1{}^t$ and $R_1{}^t$ respectively. Modify Eq. (17) by dropping asset 9 from the asset base, which gives rise to a new nominal and real rate of return on the new asset base without monetary services, which we denote by $r_{1,M}{}^t$ and $R_{1,M}{}^t$ respectively. Now modify Eq. (17) by dropping assets 8 and 9 from the asset base, which gives rise to a new nominal and real rate of return on the new asset base without inventory and monetary services, which we denote by $r_{1,IM}{}^t$ and $R_{1,IM}{}^t$ respectively. Finally modify Eq. (17) by dropping assets 5–9 from the asset base, which gives rise to a new nominal and real rate of return on

[49] Again, the last "decade" covers only the years 2010–2014.

[50] Using predicted user costs, the corresponding decade by decade geometric average rates of predicted TFP growth, $TFP_{P2}{}^t$, were as follows, with the corresponding Jorgensonian rates of growth in brackets: 2.28% (2.32), 0.39% (0.41), 0.49% (0.64), 1.35% (1.29), 1.37% (1.22) and 1.80% (1.81) per year. Note that the difference is particularly large for the 2000s.

[51] See the EUKLEMS and World KLEMS data bases on line; Jorgenson and Timmer (2016).

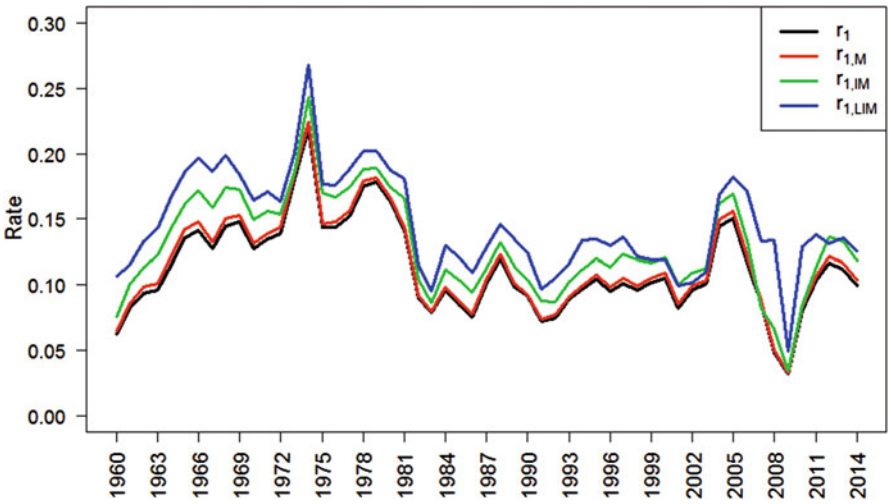


Fig. 13 Sector 1 nominal rates of return on alternative asset bases

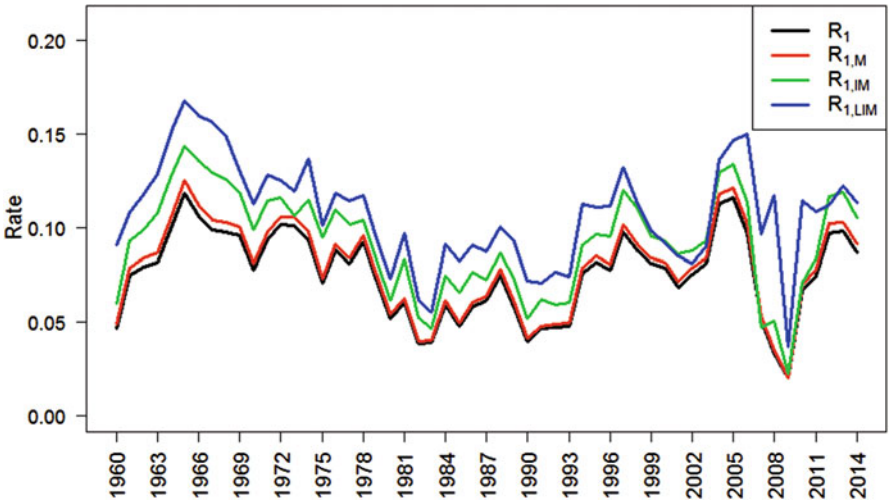


Fig. 14 Sector 1 real rates of return on alternative asset bases

the new asset base without land, inventory and monetary services, which we denote by $r_{1,\,LIM}{}^{t}$ and $R_{1,\,LIM}{}^{t}$ respectively. The alternative nominal rates can be found in Fig. 13 and the alternative real rates of return can be found in Fig. 14.

For each asset base, the value of capital services adds up to value added less the value of labour input. Thus as we decrease the number of assets in the asset base, the nominal and real rate of return on the remaining assets must increase and this fact is reflected in Figs. 13 and 14. With all assets in the asset base, the average nominal

rate of return on assets is 11.25%. Dropping monetary holdings from the asset base increases the average rate of return to 11.60% and then dropping inventory stocks further increases the average rate of return to 12.97%. Finally dropping residential, farm and commercial land from the asset base further increases the average rate of return on the remaining assets to 14.59% per year. Similarly, decreasing the asset base causes the average real rate of return on the remaining assets to go from 7.57% per year when all assets are included to 10.80% per year when money, inventories and land are dropped from the asset base. Our conclusion here is that dropping assets can substantially distort the estimated return on assets.

Recall that the year t Jorgensonian user costs $u_{1,n}{}^t$ for the nine assets used by Sector 1 were defined by Eq. (30) in Sect. 5. These user costs involved the nominal rates of return on assets for Sector 1, the $r_1{}^t$. The user costs $u_{1,n}{}^t$ were used to form the Sector 1 Jorgensonian year t capital services aggregate, $Q_{KJ1}{}^t$, and the overall Sector 1 year t input aggregate, $Q_{XJ1}{}^t$. These input aggregates along with the Sector 1 output aggregates, $Q_{VA1}{}^t$, were used to form the year t Total Factor Productivity levels, $TFP_{J1}{}^t$, for Sector 1; see Eq. (35). When we drop monetary assets from the list of assets, we obtain the new year t Jorgensonian balancing nominal rate of return for year t, $r_{1,M}{}^t$, and this new rate of return can be inserted into Eq. (30) for n = 1,...,8 in order to obtain new year t Jorgensonian user costs for Sector 1, which we define as $u_{1M,n}{}^t$. These new user costs can be used to form new year t capital services aggregates, $Q_{K1M}{}^t$, and new year t aggregate input indexes, $Q_{X1M}{}^t$, for Sector 1. In a similar fashion, when we drop both monetary assets and inventory stocks, we obtain the year t capital services aggregates, $Q_{K1IM}{}^t$, and the year t aggregate input indexes, $Q_{X1IM}{}^t$, for Sector 1. Finally, when we drop monetary assets, inventory and land stocks from the list of productive assets, we obtain the year t capital services aggregates, $Q_{K1LIM}{}^t$, and the year t aggregate input indexes, $Q_{X1LIM}{}^t$, for Sector 1. These alternative measures of aggregate capital services are used to form the alternative TFP levels, $TFP_{1M}{}^t$, $TFP_{1IM}{}^t$ and $TFP_{1LIM}{}^t$. These alternative measures of (normalized) Jorgensonian capital services and TFP are plotted in Fig. 15 along with the (normalized) measure of labour input for Sector 1, $Q_{L1}{}^t$.[52]

It can be seen that there are some small differences in the growth of Jorgensonian capital services for Sector 1 as we drop assets. With all assets included, capital services grew 6.026 fold; dropping money led to a 5.959 fold increase; dropping money and inventories led to a 5.706 fold increase and dropping money, inventories and land led to a 6.184 fold increase (see the highest line on Fig. 15). These small differences in the rates of growth of capital services as we decrease the number of assets led to even smaller differences in the rates of TFP growth. With all assets included, Jorgensonian TFP increased 2.433 fold and as we dropped assets, there were 2.444, 2.478 and 2.416 fold increases in TFP over the sample period

[52]To recover the un-normalized $Q_{L1}{}^t$, multiply the listed $Q_{L1}{}^t$ series by the value of labour input in Sector 1 for 1960, which is 180.4. To recover the four un-normalized capital services series, multiply $Q_{KJ1}{}^t$, $Q_{K1M}{}^t$, $Q_{K1IM}{}^t$ and $Q_{K1LIM}{}^t$ by the Gross Operating Surplus for Sector 1 for 1960, which is 75.5.

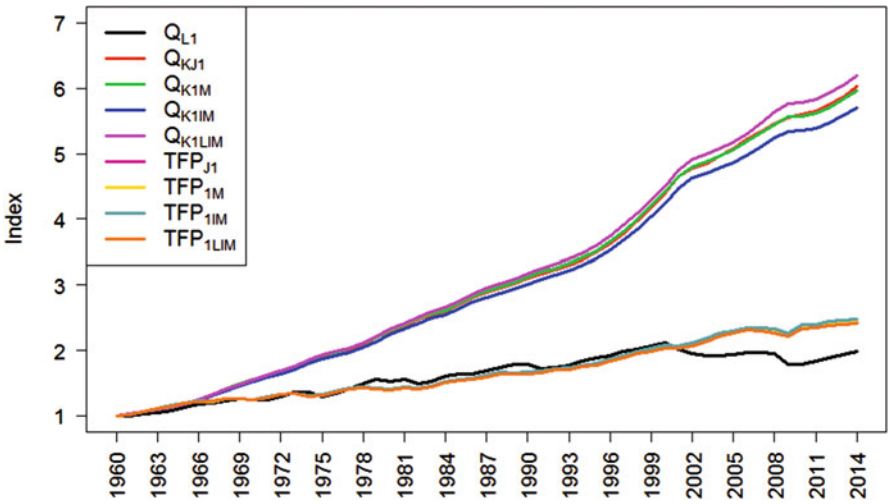


Fig. 15 Sector 1 labour and measures of capital services and TFP with alternative asset bases

for Sector 1. The corresponding geometric rates of growth were 1.661%, 1.668%, 1.695% and 1.647% per year so that all of these annual average TFP growth rates were within 0.05 of a percentage point. These differences are too small to show up in Fig. 15.

Dropping nonreproducible assets (or zero depreciation assets) from the asset base had a significant effect on ex post rates of return on assets employed in the US Corporate Nonfinancial Sector. However, dropping zero depreciation assets had a negligible effect on overall rates of TFP growth for Sector 1. In the following section, we will see if the same conclusions hold for the US Noncorporate Nonfinancial Sector.

8 Rates of Return and TFP Growth in Sector 2 with Alternative Asset Bases

Recall Eqs. (26) and (27) in Sect. 5 which defined the year t nominal and real rate of return on all fourteen assets used in Sector 2, r_2^t and R_2^t respectively. Modify Eq. (26) by dropping asset 14 from the asset base, which gives rise to a new nominal and real rate of return on the new asset base without monetary services, which we denote by $r_{2,M}^t$ and $R_{2,M}^t$ respectively. Further modify Eq. (26) by dropping assets 13 and 14 from the asset base, which gives rise to a new nominal and real rate of return on the new asset base without inventory and monetary services, which we denote by $r_{2,IM}^t$ and $R_{2,IM}^t$ respectively. Finally modify Eq. (26) by dropping assets 10–14 from the asset base, which gives rise to a new nominal and real rate of return on the new asset base without land, inventory and monetary services, which we denote

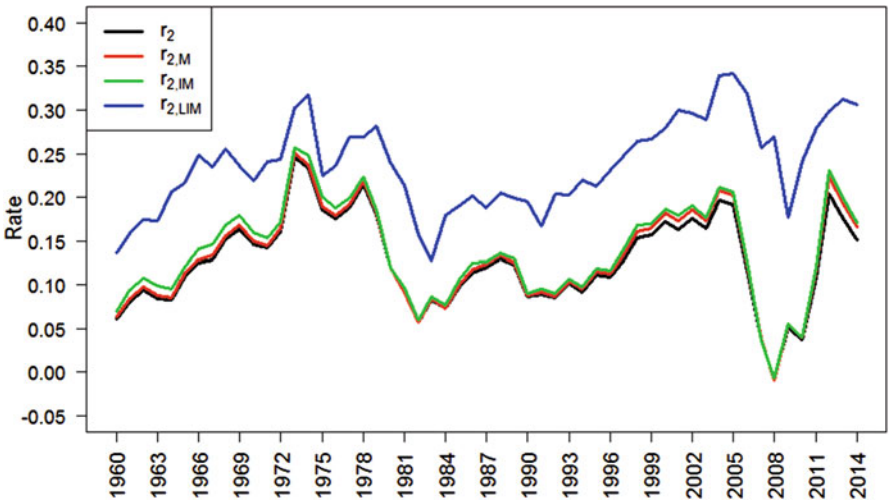


Fig. 16 Sector 2 nominal rates of return on alternative asset bases

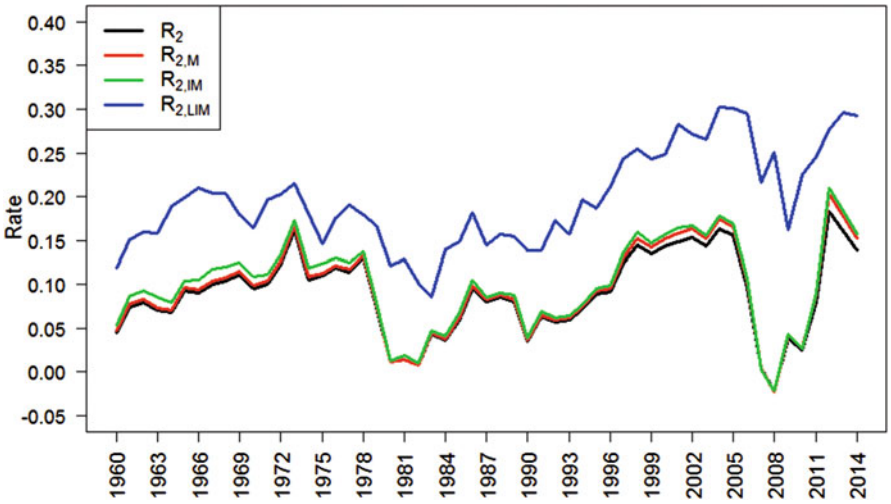


Fig. 17 Sector 2 real rates of return on alternative asset bases

by $r_{2,LIM}{}^t$ and $R_{2,LIM}{}^t$ respectively. The alternative nominal rates can be found in Fig. 16 and the alternative real rates of return can be found in Fig. 17.

It can be seen that dropping assets leads to significant increases in the measured rates of return on the asset base. With all assets included, the Sector 2 average real rate of return was 9.03%; dropping money leads to a 9.49% rate of return, further dropping inventory stocks leads to a 10.03% rate of return and further dropping land leads to a huge 19.68% average rate of return on the remaining assets. Again, our conclusion here is that dropping assets can substantially distort the estimated return on assets.

Recall that the year t Jorgensonian user costs $u_{2,n}{}^t$ for the fourteen assets used by Sector 2 were defined by Eq. (32) in Sect. 5. These user costs involved the nominal rates of return on assets for Sector 2, the $r_2{}^t$ defined by Eq. (26). The user costs $u_{2,n}{}^t$ were used to form the Sector 2 Jorgensonian year t capital services aggregate, $Q_{KJ2}{}^t$, and the overall Sector 2 year t input aggregate, $Q_{XJ2}{}^t$. These input aggregates along with the Sector 2 output aggregates, $Q_{VA2}{}^t$, were used to form the year t Total Factor Productivity levels, $TFP_{J2}{}^t$, for Sector 2; see Eq. (37). When we drop monetary assets from the list of assets, we obtain the new year t Jorgensonian balancing nominal rate of return for year t, $r_{2,M}{}^t$, and this new rate of return can be inserted into Eq. (32) for n = 1,...,13 in order to obtain new year t Jorgensonian user costs for Sector 2, which we define as $u_{2M,n}{}^t$. These new user costs can be used to form new year t capital services aggregates, $Q_{K2M}{}^t$, and new year t aggregate input indexes, $Q_{X2M}{}^t$, for Sector 2. In a similar fashion, when we drop both monetary assets and inventory stocks, we obtain the year t capital services aggregates, $Q_{K2IM}{}^t$, and the year t aggregate input indexes, $Q_{X2IM}{}^t$, for Sector 2. Finally, when we drop monetary assets, inventory and land stocks from the list of productive assets, we obtain the year t capital services aggregates, $Q_{K2LIM}{}^t$, and the year t aggregate input indexes, $Q_{X2LIM}{}^t$, for Sector 2. These alternative measures of aggregate capital services are used to form the alternative TFP levels, $TFP_{2M}{}^t$, $TFP_{2IM}{}^t$ and $TFP_{2LIM}{}^t$. These alternative measures of (normalized) Jorgensonian capital services and TFP are plotted in Fig. 18 along with the (normalized) measure of labour input for Sector 2, $Q_{L2}{}^t$.[53]

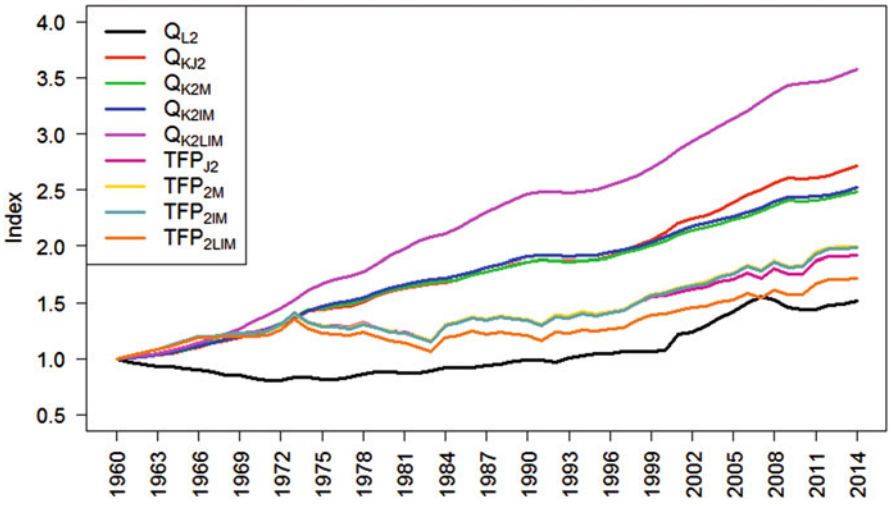


Fig. 18 Sector 2 labour and measures of capital services and TFP with alternative asset bases

[53]To recover the un-normalized $Q_{L2}{}^t$, multiply the listed $Q_{L2}{}^t$ series by the value of labour input in Sector 2 for 1960, which is 76.6. To recover the four un-normalized capital services series, multiply

It can be seen that there are some large differences in the growth of Jorgensonian capital services for Sector 2 as we drop assets. With all assets included, capital services grew 2.71 fold; dropping real monetary balances (which increased more rapidly than other assets, particularly in recent years) led to a 2.49 fold increase in the remaining capital services; dropping money and inventories led to a 2.52 fold increase and dropping money, inventories and land led to a 3.57 fold increase in the remaining capital services (see the highest line on Fig. 17). Since land stocks grow more slowly than other capital stocks and since land is a very large component of the Sector 2 capital stock, these results are not unexpected. These large differences in the rates of growth of capital services as we decrease the number of assets led to significant differences in the rates of TFP growth. With all assets included, Jorgensonian TFP increased 1.91 fold and as we dropped assets, there were 2.00, 1.99 and 1.71 fold increases in TFP over the sample period for Sector 2. The corresponding geometric average rates of TFP growth for Sector 2 were 1.21%, 1.29%, 1.28% and 1.00% per year. Thus dropping land from the list of in scope assets significantly reduced the measured rate of Jorgensonian TFP growth. Excluding money from the list of assets also had a significant (but smaller) effect.

Our conclusion is that dropping zero depreciation assets will in general significantly increase measured rates of return on assets. On the other hand, dropping zero depreciation assets will not *always* significantly affect long run average rates of TFP growth for a sector but for land intensive sectors, it is likely to significantly decrease measured long run average rates of TFP growth.

9 Changing Shares and Inequality

There has been significant recent interest in the measured fall in the labour share of income across many industrialised economies; the implication is that there has been a change in the distribution of income as households have hetereogeneous assets, and skills which are not equally substitutable with capital.[54] In this section we examine the issue of relative labour and capital shares using our two sector data set. Specifically, we consider how the shares change if we draw a distinction between value added and (net) income.

Our approach is based on that of Hayek (1941). Recall the expression of Jorgensonian user cost for asset n from either (30) or (32): $u_{m,n}{}^t \equiv [1 + r_m{}^t - (1 + i_{m,n}{}^t)(1 - \delta_{m,n}{}^t)]P_{Km,n}{}^t$, for sectors m = 1, 2. It is convenient for current purposes to express the *user cost value,* $UCV_{m,n}{}^t$, for asset n in sector m in the following form:

$Q_{KJ2}{}^t$, $Q_{K2M}{}^t$, $Q_{K2IM}{}^t$ and $Q_{K2LIM}{}^t$ by the Gross Operating Surplus for Sector 2 for 1960, which is 30.8.

[54] See, for example, Karabarbounis and Neiman (2014), Bridgman (2014) and Cho, Hwang and Schreyer (2017).

$$UCV_{m,n}{}^{t} \equiv \left[r_{m}{}^{t} - i_{m,n}{}^{t} + \left(1 + i_{m,n}{}^{t}\right)\delta_{m,n}{}^{t}\right] P_{Km,n}{}^{t} Q_{Km,n.} \tag{39}$$

Thus, for each asset n the user cost value, or capital services of asset n, can be decomposed into the sum of the following terms: financing cost, or waiting services,[55] $r_{m}{}^{t} P_{m,n}{}^{t} Q_{Km,n}$, asset revaluation, $-i_{m,n}{}^{t} P_{K,m}{}^{t} Q_{Km,n}$, and depreciation, $\delta_{m,n}{}^{t} P_{m,n}{}^{t+1} Q_{Km,n}$. Pigou (1941) argued that an appropriate measure of income is valued added less depreciation; this accounts for the physical deterioration of assets used in producing consumption goods. It is hence an income concept that emphasizes the maintenance of physical capital. Hayek (1941), however, argued that this would overstate income due to not taking into account the revaluation of assets from, for example, foreseen obsolescence. Thus, Hayek's is an income concept that emphasizes the real financial maintenance of capital.

Bridgman (2014) and Cho, Hwang and Schreyer (2017) have examined the impact on relative labour and capital shares of changing from a value added measure of income to a Pigou-type of income by subtracting depreciation from value added. Here we highlight the Hayekian concept of income, and thus also subtract asset revaluation to form our income measure.

That is, income is equal to the wage bill plus the capital stock times the ex post nominal rate of return on this stock, or $r_{m}{}^{t} P_{Km,n}{}^{t} Q_{Km,n}$ rather than the full user cost value of (39). Hence the difference between value added and this income measure is the value of depreciation and asset revaluation.

A comparison of nominal value added with Hayekian and Pigouvian nominal income is provided in Fig. 19 for Sector 1.[56] It can be seen that nominal value added is generally higher than nominal income, especially since 2007. Comparing the Hayekian and Pigouvian income measures, it can be seen that the Hayekian measure is typically larger, due to positive asset revaluations in most years, and more volatile. With depreciation rates evolving relatively smoothly,[57] changes in prices of residential and commercial land in particular appear to drive much of this difference in volatility, especially around 2008.[58]

The share of capital services in value added is the user cost value of (39) summed over all assets and divided by nominal value added. These shares are plotted in Fig. 20.[59] The greater volatility of nominal Hayekian income seen in Fig. 19 is reflected in the capial income shares in Fig. 20.[60] The generally lower capital shares in either Hayekian of Pigouvian income indicate less inequality than implied by the corresponding value added shares. In terms of long-term trends, the share of capital services in value added goes from 0.295 in 1960 to 0.367 in 2014 (a 24% increase), while our preferred share of capital in total Hayekian income goes from

[55] See Rymes (1969, 1983) on the concept of waiting services.

[56] Jorgensonian ex post rates of return are appropriate in this context; see Sect. 5.

[57] See Appendix A8 of Diewert and Fox (2016).

[58] See tables A1 and A9 of Diewert and Fox (2016).

[59] The value added shares are the same as those in Table 1.

[60] Labour shares are of course a mirror image of these capital shares.

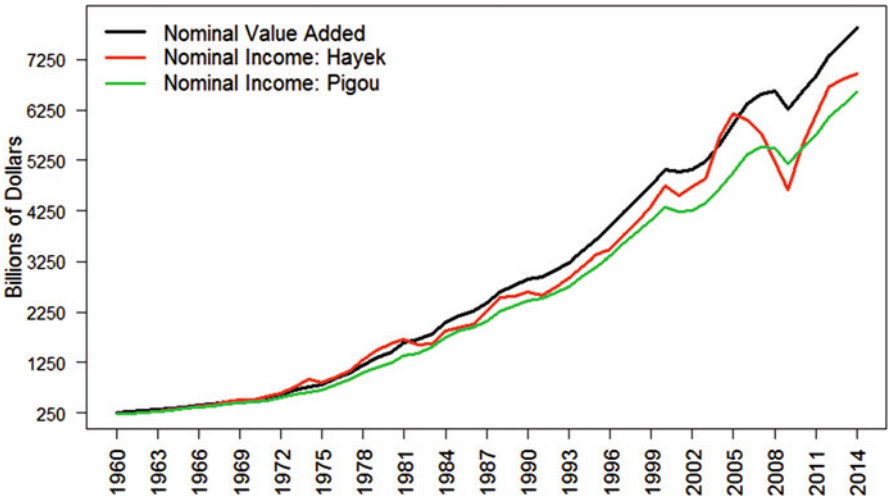


Fig. 19 Sector 1 nominal value added and nominal income

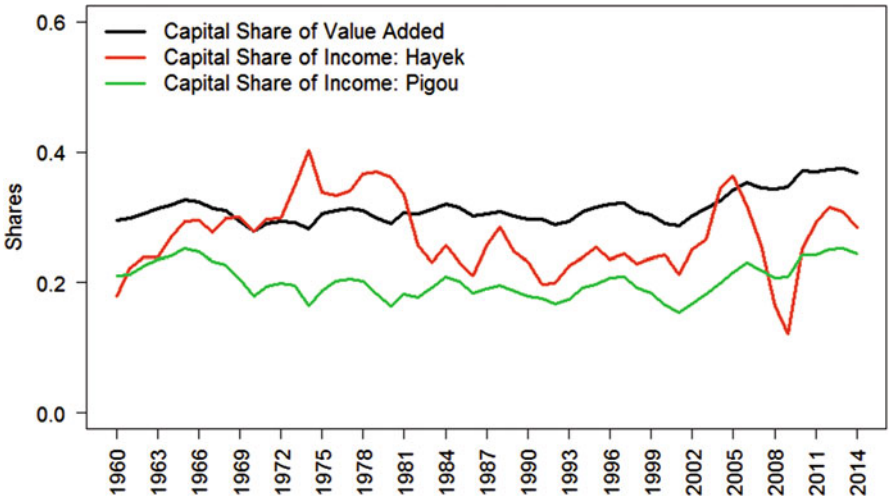


Fig. 20 Sector 1 capital shares of value added and income

0.179 in 1960 to 0.283 in 2014 (a 58% increase). Thus, while all capital shares have grown, the Hayekian income share has grown more, although with much higher year-on-year volatility. This somewhat strengthens the view of long-term increasing inequality through a shift in the relative distribution of income from labour to capital.

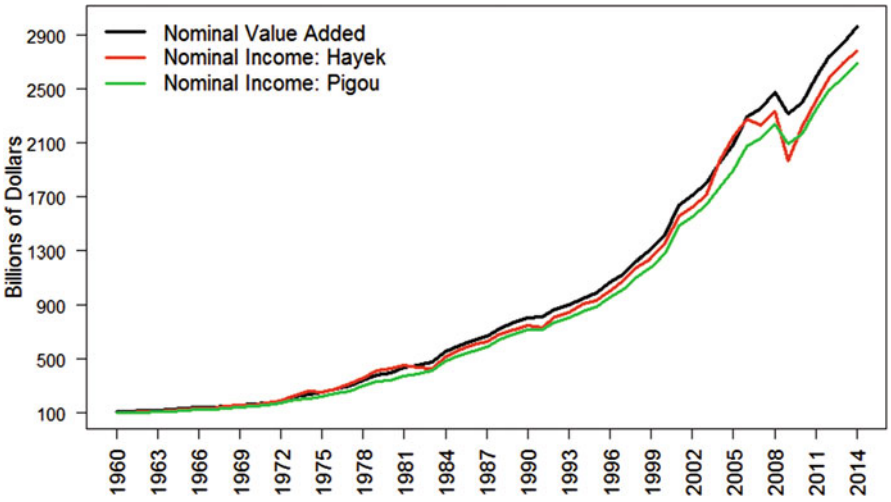


Fig. 21 Sector 2 nominal value added and nominal income

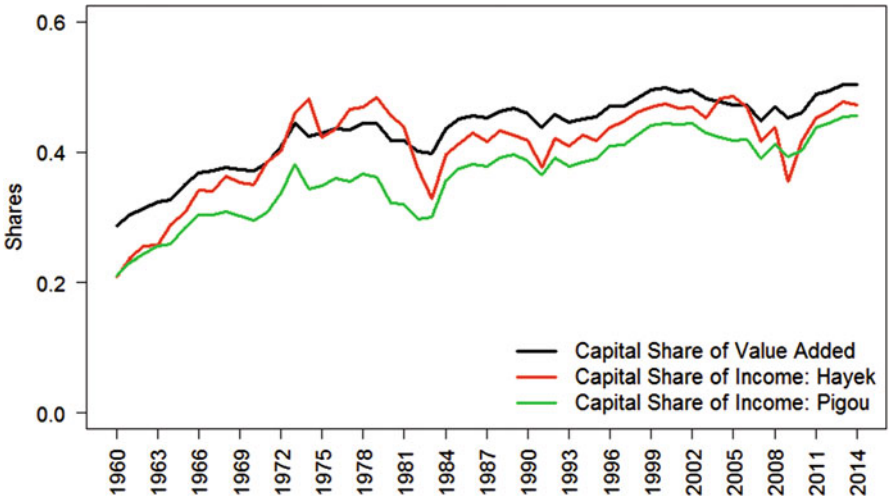


Fig. 22 Sector 2 capital shares of value added and income

The corresponding results for Sector 2 are shown in Figs. 21 and 22.[61] From Fig. 21 we again see the increased deviation between value added and income from 2007. As for Sector 1, the shares in Fig. 22 indicate less inequality using income shares compared to value added shares. In terms of long-term trends, the share of

[61] The value added shares are the same as those in Table 2.

capital services in value added goes from 0.287 in 1960 to 0.504 in 2014 (a 76% increase), while the share of capital income in total Hayekian income goes from 0.209 in 1960 to 0.472 in 2014 (a 125% increase). Hence, while all capital income shares have grown, the Hayekian income share has grown more. Thus, evidence for Sector 2 also strengthens the view of increasing inequality through a shift in the relative distribution of income from labour to capital.

10 Conclusions

A number of tentative conclusions can be drawn from the above analysis:

1. The technologies used in the Corporate Nonfinancial (Sector 1) and Non-corporate Nonfinancial (Sector 2) sectors are quite different. Sector 1 uses reproducible assets quite intensively while Sector 2 uses land and structure assets quite intensively.
2. Total Factor Productivity growth in our two sectors over the years 1960–2014 has been excellent: the TFP growth rate for Sector 1 averaged 1.66% per year using Jorgensonian user costs (1.69% per year using predicted user costs) and 1.21% per year for Sector 2 using both sets of user costs. These are very high average rates of TFP growth over such a long period, even though there has been a significant productivity slowdown over 2005–2014, especially for Sector 1.
3. Average real rates of return on productive assets employed have been quite high in both sectors. The average annual real rate of return was 7.6% per year in Sector 1 and 9.0% per year in Sector 2. There is no indication of a long run slowdown in these rates of return (but there have been massive short run fluctuations in these rates).
4. Jorgensonian user costs use actual ex post asset inflation rates in place of predicted asset inflation rates and as a result, Jorgensonian user costs are volatile and frequently negative if land assets are included in the asset boundary. These user costs are not suitable for many analytical purposes. Our predicted asset inflation rates generated relatively smooth user costs that could be used in production and cost function studies. However, Jorgensonian user costs are the right type of user cost to use when calculating ex post rates of return on assets employed.
5. Somewhat surprisingly, Jorgensonian and predicted user costs can give rise to rates of growth of capital services and Total Factor Productivity that are very close to each other. Thus for Sector 1, we found that the long run average geometric rate of capital services growth generated by the alternative user cost approaches were 3.38% and 3.29% per year for Sector 1 and 1.86% and 1.88% per year for Sector 2. The resulting alternative annual rates of TFP growth were 1.66% and 1.69% per year in Sector 1 and 1.21% per year using both Jorgensonian and predicted user costs for Sector 2. These differences are not large.

6. Dropping assets from the asset base can lead to very large biases in the measured rates of return on assets employed. Dropping land, inventory and monetary balances from the list of assets in scope increased the measured average ex post real rate of return on assets from 7.6% to 10.8% per year for Sector 1 and from 9.0% to 19.7% per year for Sector 2.
7. Dropping assets from the asset base can lead to little change in measured TFP growth rates or it can lead to significant changes. Thus the Jorgensonian average TFP growth rate for Sector 1 changed from 1.66% per year with all assets in the base to 1.65% per year, after land, inventories and real monetary balances were dropped from the list of assets. On the other hand, the Jorgensonian average TFP growth rate for Sector 2 changed from 1.21% per year with all assets in the base to 1.00% per year after land, inventories and real monetary balances were dropped from the list of assets. This is a significant change.
8. Our data are subject to a considerable degree of uncertainty. Hopefully, in future years, the BEA in cooperation with the BLS, the USDA and the Federal Reserve Board of Governors will be able to improve the quality of the underlying data. In particular, we note that our land and labour data are weak and we are missing data on resource stocks.
9. More research is needed on choosing appropriate predicted asset inflation rates.
10. Using (net of depreciation and asset revaluation) income is more appropriate than value added for examining changes in the relative distribution of income between labour and capital, with potentially different results relating to the extent of changes in inequality.

References

Babbage, C. (1835). *On the economy of machinery and manufactures* (4th ed.). London: Charles Knight.

Böhm-Bawerk, E. V. (1891). *The positive theory of capital*, W. Smart (translator of the original German book published in 1888). New York: G.E. Stechert.

Bridgman, B. (2014). "Is labor's loss capital's gain? Gross versus net labor shares", manuscript, Bureau of Economic Analysis, Washington D.C.

Cho, T., Hwang, S., & Schreyer, P. (2017). Has the Labour Share Declined?: It Depends, *OECD Statistics Working Papers*, 2017/01. Paris: OECD Publishing.

Christensen, L. R., & Jorgenson, D. W. (1969). The measurement of U.S. Real capital input, 1929–1967. *Review of Income and Wealth, 15*, 293–320.

Church, A. H. (1901). The proper distribution of establishment charges, parts I, II, and III. *The Engineering Magazine, 21*, 508–517; 725–734; 904–912.

Dean, E., & Harper, M. (2001). The BLS productivity measurement program. In C. R. Hulten, E. R. Dean, & M. J. Harper (Eds.), *New developments in productivity analysis* (Vol. NBER Studies in Income and Wealth Volume 63, pp. 55–84). Chicago: University of Chicago Press.

Diewert, W. E. (1974). Intertemporal consumer theory and the demand for durables. *Econometrica, 42*, 497–516. Dordrecht, Holland

Diewert, W. E. (1976). Exact and Superlative Index Numbers. *Journal of Econometrics, 4*, 114–145.

Diewert, W. E. (1977). Walras' theory of capital formation and the existence of a temporary equilibrium. In E. Schwödiauer (Ed.), *Equilibrium and disequilibrium in economic theory* (pp. 73–126). Reidel Publishing Co.

Diewert, W. E. (1980). Aggregation problems in the measurement of capital. In D. Usher (Ed.), *The measurement of capital* (pp. 433–528). Chicago: The University of Chicago Press.

Diewert, W. E. (1992). The measurement of productivity. *Bulletin of Economic Research, 44*, 165–198.

Diewert, W. E. (2005a). Issues in the measurement of capital services, depreciation, asset price changes and interest rate. In C. Corrado, J. Haltiwanger, & D. Sichel (Eds.), *Measuring capital in the new economy* (pp. 479–542). Chicago: University of Chicago Press.

Diewert, W. E. (2005b). On measuring inventory change in current and constant dollars, Discussion Paper 05–12, Department of Economics, The University of British Columbia, Vancouver, Canada, V6T 1Z1.

Diewert, W. E. (2010). User costs versus waiting services and depreciation in a model of production. *Journal of Economics and Statistics, 230*(6), 759–771.

Diewert, W. E. (2014a). The treatment of financial transactions in the SNA: A user cost approach. *Eurostat Review of National Accounts and Macroeconomic Indicators, 1*, 73–89.

Diewert, W. E. (2014b). US TFP growth and the contribution of changes in export and import prices to real income growth. *Journal of Productivity Analysis, 41*, 19–39.

Diewert, W. E., & Fox, K. J. (2015). *Money and the measurement of total factor productivity*, Paper presented at the IARIW-OECD Special Conference: "W(h)ither the SNA?", Paris, France, April 16–17.

Diewert, W. E., & Fox, K. J. (2016). Alternative user costs, rates of return and TFP growth rates for the US Nonfinancial Corporate and Noncorporate Business Sectors: 1960–2014.

Diewert, W. E., & Fox, K. J. (2017). Decomposing value added growth into explanatory factors. In E. Grifell-Tatjé, C. A. K. Lovell, & R. Sickles (Eds.), *The Oxford handbook of productivity analysis*. Oxford University Press, New York, USA.

Diewert, W. E., & Morrison, C. J. (1986). Adjusting output and productivity indexes for changes in the terms of trade. *Economic Journal, 96*, 659–679.

Edwards, E. O., & Bell, P. W. (1961). *The theory and measurement of business income*. Berkeley: University of California Press.

Eurostat, International Monetary Fund, OECD, United Nations and World Bank. (1993). *System of national accounts 1993*. Brussels/Luxembourg/New York/Paris/Washington DC.

Eurostat, International Monetary Fund, OECD, United Nations and World Bank. (2008). *System of national accounts 2008*. Luxembourg/New York/Paris/Washington DC.

Fischer, S. (1974). Money and the production function. *Economic lnquiry, 12*, 517–533.

Harberger, A. C. (1998). A vision of the growth process. *American Economic Review, 88*(1), 1–32.

Harper, M. J., Berndt, E. R., & Wood, D. O. (1989). Rates of return and capital aggregation using alternative rental prices. In D. W. Jorgenson & R. Landau (Eds.), *Technology and capital formation* (pp. 331–372). Cambridge MA: The MIT Press.

Hicks, J. R. (1961). The measurement of Capital in Relation to the measurement of other economic aggregates. In F. A. Lutz & D. C. Hague (Eds.), *The theory of capital* (pp. 18–31). London: Macmillan.

Hill, R. J., & Hill, T. P. (2003). Expectations, capital gains and income. *Economic Inquiry, 41*, 607–619.

Jorgenson, D. W. (1989). Capital as a factor of production. In D. W. Jorgenson & R. Landau (Eds.), *Technology and capital formation* (pp. 1–35). Cambridge MA: The MIT Press.

Jorgenson, D. W. (1995). *Productivity: Volume 1, postwar U.S. economic growth*. Cambridge, MA: The MIT Press.

Jorgenson, D. W. (1996). *Investment: Volume 2; tax policy and the cost of capital*. Cambridge, MA: The MIT Press.

Jorgenson, D. W., & Griliches, Z. (1967). The explanation of productivity change. *The Review of Economic Studies, 34*, 249–283.

Jorgenson, D. W., & Griliches, Z. (1972). Issues in growth accounting: A reply to Edward F. Denison. *Survey of Current Business, 52*, 4. Part II (May), 65–94.
Jorgenson, D. W., & Nishimizu, M. (1982). U.S. and Japanese economic growth, 1952–1974: An international comparison. *Economic Journal, 92*, 707–726.
Jorgenson, D. W., & Timmer, M. (2016). *World KLEMS*, Harvard University and the University of Groningen. http://www.worldklems.net/index.htm.
Karabarbounis, L., & Neiman, B. (2014). The global decline of the labor share. *Quarterly Journal of Economics, 129*, 61–103.
Kohli, U. (1990). Growth accounting in the open economy: Parametric and nonparametric estimates. *Journal of Economic and Social Measurement, 16*, 125–136.
Peasnell, K. V. (1981). On captial budgeting and income measurment. *ABACUS, 17*, 52–67.
Pigou, A. C. (1941). Maintaining capital intact. *Economica, 8*, 271–275.
Rymes, T. K. (1969). Professor read and the measurement of Total factor productivity. *Canadian Journal of Economics, 1*, 359–367.
Rymes, T. K. (1983). More on the measurement of Total factor productivity. *Review of Income and Wealth, 29*, 297–316.
Schreyer, P. (2009). *Measuring capital - OECD manual 2009* (2nd ed.). Paris: OECD.
Schreyer, P. (2012). Measuring multifactor productivity when rates of return are endogenous. In W. E. Diewert, B. M. Balk, D. Fixler, K. J. Fox, & A. O. Nakamura (Eds.), *Price and productivity measurement: Volume 6: Index number theory* (pp. 13–40). Trafford Publishing, Vancouver, Canada. www.trafford.com.
Schreyer, P., Diewert, W. E., & Harrison, A. (2005). *Cost of capital services and the national accounts*, presented at the July 2005 Meeting of the Advisory Expert Group on the Update of the 1993 SNA, Bangkok. http://unstats.un.org/unsd/nationalaccount/aeg/m1-05.asp.
Theil, H. (1967). *Economics and information theory*. Amsterdam: North-Holland.
Törnqvist, L. (1936). The Bank of Finland's consumption price index. *Bank of Finland Monthly Bulletin, 10*, 1–8.
Törnqvist, L., & Törnqvist, E. (1937), Vilket är förhällandet mellan finska markens och svenska kronans köpkraft?", *Ekonomiska Samfundets Tidskrift* 39, 1–39 reprinted as pp. 121–160 in *Collected Scientific Papers of Leo Törnqvist*, Helsinki: The Research Institute of the Finnish Economy, 1981.
von Hayek, F. A. (1941). Maintaining capital intact: A reply. *Economica, 8*, 276–280.
von Neumann, J. (1945). A Model of General Economic Equilibrium. *Review of Economic Studies, 12*, 1–9.
Walras, L. (1954). *Elements of Pure Economics*, a translation by W. Jaffé of the Edition Définitive (1926) of the *Eléments d'économie pure*, first edition published in 1874, Homewood, Illinois: Richard D. Irwin.

On the Allocation of Productivity Growth and the Determinants of U. S. Income Inequality

Shasha Liu, Robin C. Sickles, and Shiyi Zhang

Abstract We estimate the determinants of income inequality focusing on the U.S. over 1985—2013. We decompose widely used inequality indices by subgroups, income sources, and factor components based on regression approaches to discover the significant sources of widening inequality in the United States. Our results indicate that education, marriage, gender, race, asset income, as well as employment in manufacturing and financial industries all expand the gap across different income classes.

Keywords Decomposition methods · Income inequality · Entropy indices · Demographics

JEL Classification: D31, J10, J31 (Index)

1 Introduction

Productivity and efficiency research has been the focus of an expanding body of literature in the last few decades. With increased total factor productivity the world economy has seen continued economic growth and more integrated global

This paper is based on the Keynote Address by Robin Sickles to the North American Productivity Workshop IX, Quebec City, Canada, June 15-18, 2016, "Total Factor Productivity: Secular Movements and Income Inequality," and on the May 5, 2016 Rice University Senior Honors Thesis of Shiyi Zhang, *Determinants of Income Inequality in the U.S: Decomposition Methods*.

S. Liu • R.C. Sickles (✉)
Department of Economics, Rice University, Houston, TX, USA
e-mail: shasha.liu@rice.edu; rsickles@rice.edu

S. Zhang
Department of Economics, Department of Computational and Applied Mathematics, Rice University, Houston, TX, USA
e-mail: sz26@rice.edu

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_3

supply chain network. However, economic challenges exist and feasible solutions to them appear far from certain. One of the deep and important issues of our time is the distributional implication of productivity growth and the closely related issue of rising income inequality. The gap between the growth in productivity and wage compensation has widened since the late 1970s in the United States, and it has been a phenomenon across the world. This poses important research questions as to why inequality has risen and why relatively recent productivity growth disproportionately has been allocated to capital returns and away from wage compensation.

Looking at how wage compensation has changed in the last three decades, one is alarmed by the stagnating median wage in the U.S. compared to significant productivity growth and rising inequality indices. In fact, the U.S. has become the most unequal country among developed countries as well as one of the unequal countries in the world in terms of wealth and income, with a Gini index based on income increasing by over 30% from 1979 to 2013 (0.31–0.40). The bottom 90% of households has seen their incomes stagnate over the last three decades. Against this backdrop of increasing income inequality is the misconception by the voting public that income inequality is not as severe as it is and that free-market capitalism is the working mechanism to ameliorate the problem with no need for public policies to assist in this process.

What happened that prevented the growth in wage compensation from keeping up with the growth in TFP? We will discuss different socioeconomic aspects that may have contributed to the gap between the growth in productivity and wage compensation.

Technology is arguably one of the most important factors contributing to the gap between wage and productivity growth. Many believe that technological innovation is the key factor that has increased income inequality since the late 1970s. The U.S. is in transition from manufacturing to a service economy with significant technological advances that have improved TFP over the last three decades. The transition has created jobs that require highly skilled workers to perform tasks such as writing software, and service jobs such as customized tailoring. However, jobs in the middle which compose the majority of manufacturing jobs are shrinking. According to Brynjolfsson and McAfee (2011), the demand for some types of skills or abilities will likely decrease significantly with technological advancements, worsening the employment levels and wages of people with those skills or work in related industries. In addition, the recent recession accelerated the displacement of these mid-wage jobs, according to a study by Cortes and his colleagues (2014). People blame computer technology for the lost jobs. Indeed, computers are more likely to replace routine jobs, leaving those with no advanced training trapped in unemployment. Tyler Cowen (2013) argued that computing technology accelerates the division of American workforce into two classes: the highly skilled workforce that will tremendously benefit from working with technologies, and the other group whose wages will stagnate or even decline. Autor, Katz and Kearney (2008) further pointed out that inequality within the upper half of the income class keeps increasing, while the inequality trend in the lower half of the distribution is slowing

down. On the other hand, it is still debatable whether technical innovation is the main source of income inequality if we account for the complementary effect of information technology (IT) on the labor market. Certain technologies create professions for which creativity and dexterity are indispensable and thus boost demand for certain types of labor. Skilled manual labor, such as sewing machine operators, boilermakers, etc., is currently fortunate professions that have had the largest recent wage gains. In addition to replacing routine jobs, technology also impacts non-routine jobs which were considered to be confined to humans. The thought that machinery will replace all these jobs challenges our perspective on technology. As Moshe Vardi (2016) pointed out, what would technology leave us in the future with all these jobs replaced by machines and how would the increased productivity from the IT revolution benefit social wellbeing?

Even though technology may be more easily linked to the current dynamic of capital deepening, education is considered by many to be the underlying source of income inequality. As technological innovation has accelerated, increasing educational attainment is necessary to earn higher incomes. Goldin and Katz (2008) have noted that the U.S. economy bloomed during the twentieth century largely because the educational attainment of the U.S. population raced along with technological improvements. However, educational attainment has not kept up with recent changes in technology. The educational choices Americans make will determine if they can gain higher wages in the computing era and shape the future of the U.S. economy. Not everyone holds the same view that we should increase educational attainment to keep up with technological development, however. Autor (2015) has noted that highly skilled workers have less rapid career growth and are moving into less skilled occupations, and that technology may not be progressing as rapidly as people once thought. Productivity growth has slowed down dramatically after 1995. A significant employment growth in the past 15 years has been in the relatively low-education, in-person service occupations. As Paul Krugman has also pointed out in a *New York Times* article on February 23, 2015, educational failing is not the reason for income inequality. Employers offer higher wages to attract workers with certain skills to perform tasks that require visual perception, mobility, and etc.

The significant income growth at the top of the income distribution over the last three decades has been the driving force in productivity growth and in also making the U.S. income distribution one of the most unequal. According to Sabadish and Mishel (2012), annual earnings of the top 0.1% have increased by more than 300% since the late 1970s, while the bottom 10% of earners only have seen a 34% increase in their wages. The share of total income in the U.S. received by the top 1% of earners has doubled from the late 1970s to pre-recession. An important contributor to the drastic difference in income growth is CEO compensation and the financial industry. Average executive compensation has increased more than 725%, much greater than stock market growth, while average worker compensation has fallen by 5.7% from 1979 to 2005.

The increasing presence of "power couples" may also play an important role in driving income inequality. Couples are more likely to have similar backgrounds such as educational attainment, earnings potential, and lifestyle. The fact that

women have had increasing educational attainment as well as growing participation in the labor force has changed their social role and status in marriage. Women have more freedom in choosing their partners and tend to marry those who share similar backgrounds. This assortative mating process leads to higher levels of social segregation and inequality, and the impact is magnified through future generations who are likely to share their parents' values and conventions. Inequality thus reinforces itself by increasing the incentive for assortative mating.

Additional factors that have significant impacts on income inequality are race and gender. The income gap between different races, particularly between white and black families, has continued even as the U.S. is becoming more diverse. The income disparity between white and minority groups aggravates problems with political and economic instability. Moreover, the reduction in the progressivity of the federal tax system in the early 1980's has reinforced inequality by increasing the after tax incomes and the wealth of those with high incomes and wealth. Although women have more access to employment opportunities in today's economy than in years past, the gender wage gap still exists even though women receive more college and graduate degrees than men and have substantially increased their share of employment. Wage differentials still exist for the same job for those with comparable skill and experience levels in almost every occupation, although the exact percentage gap is subject to some debate. However, a consensus would agree that persistent occupational segregation is a contributor to the lack of significant progress in closing the wage gap. Outright racial and gender discrimination in pay, hiring, or promotions is still a significant feature in workplaces.

We study these potential causes of income inequality and how they impact income distributions. Analyzing the role of inequality that emanates from the dynamic of productivity growth is essential if one is to understand and address its social and economic challenges. Labor displacement, healthcare reform, and political stability are still major social and economic issues and these issues are fundamentally linked to inequality. Being able to answer why inequality has risen and how inequality is connected to productivity growth will better inform us of the implication of polices meant to remedy this dynamic.

2 Literature Review

2.1 Technology

Over the past several decades, technological advances, especially computing ones, have improved the total factor productivity of the U.S. (Cardarelli and Lusinyan 2015). IT-intensive industries have experienced a larger acceleration in productivity than other industries and have been the most important contributor to this productivity growth (Stiroh 2002). Nevertheless, despite the substantial growth of productivity since the 1970s, the median American male worker's real wage rose by only 3% from the late 1970s to 2014, as pointed out by many researchers

(DeNavas-Walt and Proctor 2014). Technology confers tremendous benefits on non-wage earners and may have led to income growth among all income groups (Mishel et al. 2012).

Goldin and Katz (2008) have pointed out that technological innovation, especially the involvement of the computer in the workplace, is the key factor that gave rise to the increase in income inequality between 1979 and 2008. They argued that new technologies alter the relative demand for different types of labor by shifting the demand for skilled workers, thus changing the overall wage structure. Cowen (2013) has proposed that the acceleration of computing technological innovation will result in two classes in America: a highly skilled class that will profit tremendously by learning to work with technologies, and the rest of the population whose wages will stagnate or even decline. Lawrence Summers has pointed out in a *Wall Street Journal* article on July 7, 2014 that technological advance is the driving factor in the reducing job opportunities in many industries.

According to Autor, Katz, and Kearney (2008), the relation between changing demands for job tasks and the pervasiveness of computerization also can partially to explain 'polarization'—inequality within the upper half of the income class keeps increasing, while the inequality trend in the lower half of distribution gradually slows down. IT is quick to substitute middle skills for occupations requiring routine tasks, such as certain bank clerks' work that has already been replaced by the ATM technology. Computerization, on the other hand, cannot substitute the jobs that require more 'abstract tasks' and cognitive and interpersonal skills, such as lawyers, managers, and professors, at least not yet. Furthermore, some non-routine 'low-skilled' services jobs, like security guards, health aids, and servers, are also hard for technologies to replace. Autor, et al. concluded that due to the substitution effect, technological advance increases labor demand for the 'two tails' of skilled workers, but reduces demand for 'middle skilled' labor.

However, both Summers and Autor later renounced their previous assertion of technology as the major driver of inequality after taking into account of the complementary effect of IT on the labor market. Specifically, certain technologies such as information process and motor power are prone to create professions, to which creativity and dexterity are indispensable, thereby boosting demand for labor. Moreover, despite the general consensus on the pervasive effect of technology on the labor market, there is not yet a consensus, pointed out by Summers, on the overall effect of technology as a substitute or a complement to labor and thus wage earners and thus whether technological innovation is a leading main source of income inequality in the US.

2.2 Executive Compensation and Financial-Sector Pay

Sabadish and Mishel (2012) have noted that the wage gap between the very highest earners and other earners, including low, middle and other high earners, might be one of the sources that contribute to overall income inequality in the United

States. They pointed to the remarkable growth of earnings at the top of the income distribution as a major cause of income inequality. Their study shows that the annual earnings of the top 0.1% has grown 362% since the late 1970s, approximately ten times than the 34% of wage growth in the 90th to 95th percentile. Additionally, from 1979 to 2007, the share of total income in the United States received by the top 1% of earners doubled.

Sabadish and Mishel ascribed income inequality to CEO compensation and growth in the financial industry. They found that a large number of the highest earning households are headed by either executives or those employed in the financial sector. These professions account for 58% of income expansion in the top 1% and 67% in the top 0.1% from 1979 to 2005.

2.3 Workplace Heterogeneity

Few studies have been conducted to investigate how income levels differ along the lines of firm types and worker allocation across those firms. We now turn to research that has focused on the inequality issue from the workplaces' perspective.

Card, Heining and Kline (2013) examined the effects of workplace heterogeneity on wage inequality. Even though their empirical study utilized German data, it offered potential explanations for inequality in the United States. Card et al. examined how much of the rise in inequality could be attributed to a rise in the variation in pay premiums offered by different employers, and on how much could be explained by rising heterogeneity among workers. By decomposing changing inequality patterns in this way, they showed that both factors partly contribute to overall inequality. Along with the growth in the gap between consistently high- and low-wage workers is the expanding divergence between the high- and low-pay jobs, which suggests the increasing importance of the often education-mediated job search and matching process. Card, et al.'s results show that people with higher pay increasingly cluster at establishments that generally pay higher wages to their employees. This polarization causes inequality in average wages across industries to rise substantially. Card et al. also found that the distribution of establishment effects is relatively disperse for new establishments and thus concluded that new establishments contribute more to the workplace heterogeneity effects on wage inequality. Barth and his colleagues (2016) applied Card's approach to study inequality in the U.S. and illustrated that workplace heterogeneity is an important part of recent rises in wage inequality in the United States.

2.4 Education

Considered an indispensable part of economic growth, disparate educational opportunities, along with technology effects, have been singled out by many economists

as a leading cause of income inequality. Goldin and Katz (2008) called the twentieth century 'The American Century', as well as the 'Human Capital Century', indicating that economic growth in the U.S. in the past century was largely fueled by the increasing educational attainment of the U.S. population. Kearney, Hershbein and Boddy (2015) from the Hamilton Project also pointed out that education is integral to inequality, as more advanced education is necessary for laborers to capture higher wages in job market. Card, Heining and Kline (2013) also noted that higher productivity normally associated with higher education tends to make higher degree holders more attractive to the higher-paying establishments, leading to higher overall inequality.

Before the 1970s, the increasing popularity of high-school degree attainment rendered real income growth fastest near the bottom of income distribution and slowest at the top, slightly decreasing overall income inequalities (Goldin and Katz 2008). However, after the 1970s, family incomes grew much faster in the top 5% income-class than the middle and the lowest quintile. Goldin and Katz explained this phenomenon by pointing to a race between education and technological advancement. Before the 1970s, there was an increasing supply of educated American to meet the increasing demand for skilled workers. Hence, inequality did not rise dramatically, and economic gains were broadly shared. But afterward, as accelerated technological innovation has outpaced the growth of educational attainment and gains from technology have increased, people equipped with technological skills from higher education tend to earn much higher incomes.

While some researchers emphasized the failure of an increase in educational attainment to keep up with technological development, Autor (2015) argued that it is not the main reason for wage inequality. He argued that technological progress is not growing as rapidly as people thought. He observed that the premium to higher education has stagnated over the last 10 years, as Paul Krugman in his *New York Times* article on February 23, 2015 also argued that inflation-adjusted earnings of highly educated Americans have plateaued since the 1990s. On the other hand, there is employment growth in the relatively low-education, in-person service occupations, which offer higher wages to attract workers who typically don't receive high education. For example, as we pointed out earlier, sewing machine operators and boilermakers have seen wage gains recently.

2.5 Race

It is obvious that income in the United States is not distributed evenly among races. Mishel and his colleagues (2012) showed that family money income differs significantly across racial and ethnic groups. In 2010, the median income of black families and that of Hispanic families was $39,715 and $40,785, which were less than 63% of the median of white families income $65,138.

Wright (1978) showed that it is much harder for blacks to step into high-income class than non-blacks and this appears not to have changed substantially

in the ensuing almost 40 years. Both race and the present racial discrimination play a critical role in income inequality across racial lines. First, racial identity can be a useful dividing line that often corresponds to wage levels, as the current prevailing criteria built into the mechanisms for both the job sorting and the promoting process tend to favor white people while bringing disadvantages to the black people. However, rather than as a cause for the wage gap, race simply helps explain this phenomenon, since many criteria such as educational credentials and connections themselves are not racist in nature. On the other hand, the still existing pattern of racial discrimination and the social dynamics of domination in managerial hierarchies may tend to legitimatize the current racially informed job position hierarchy. As a result, Wright claimed that blacks are more likely to cluster at the working class level and have little chance to gain promotion into managerial positions, in which they would gain much higher income. These factors appear no less relevant today than they did 40 years ago.

2.6 Gender

The report launched by the Council of Economic Advisers of the White House (2015) suggests income inequality along gender lines, despite the apparent symmetry of female and male worker numbers. This study contrasts with the previous data, which shows that the gender wage gap has generally narrowed since 1970s, by pointing out the still existing difference—in 2013, the median full-time working females only earned 78% of what earned by median male full-time workers. Moreover, beyond wages, employer-sponsored health and retirement benefits and other compensation should also be taken into account when considering the causes for income difference between genders. Study shows that it is less likely for women, especially those women with lower incomes, to have retirement saving plans.

Furthermore, even though women take almost 50% of job positions, women tend to cluster at lower-paying occupations and industries. Representing more than half of employers in the three industries with the lowest average wages—leisure and hospitality, retail trade, and other services—women continue to have relatively low shares in the three industries with the highest average wages—information services, mining and logging, and utilities. Therefore, the gender income gap is partly an occupational one as shown in the allocation pattern of male and female across industries.

2.7 Marriage

Studies have found that the patterns of marriage also affect income inequality. Homogamy within households has increased more for the most and least educated households than for those in the middle (Kalmijn 1991, Rosenfeld 2008, Schwartz

and Mare 2005). This explains that the real earnings of households with the least education decline while those of households headed by high school graduates stagnate for years. The findings suggest that educational homogamy leads to the increasing income gap between high and low income couples. Schwartz (2010) also finds the effects of the growing earning homogamy between spouses on inequality across married-couple families to be 17–51% depending on the measures and time periods. The earning homogamy can be explained partially by the increase in positive assortative mating.

Positive assortative mating is termed by sociologists to describe the phenomenon that people search to marry those with similar backgrounds including education attainment, earning potential, value systems, personal preferences, and lifestyle. Studies have found that assortative mating increases household income inequality at all levels of education (Eika et al. 2014). The time trend is particularly evident for less educated who are more likely to sort into homogamous marriages. Further finding suggests that educational assortative mating contributes significantly to the income inequality of households with different educational backgrounds. Consequently, we speculate that the assortative process is magnified through replication into the future generations. More likely to be raised by couples with the same backgrounds, children tend to resemble their parents' level of income, as their opportunities and achievements are highly correlated with their parents' education and income level.

3 Decomposition Methods

Many economists have utilized income inequality decompositional analyses to study and quantify the effects of various factors on income inequality. Shorrocks (1983) used data on the distribution of net family incomes in the United States between 1968 and 1977 to determine the proportion of total income inequality attributed to various factors. Heshmati (2004) discussed methods to decompose income inequality, including through subgroups, income sources, and factor components such as characteristics of households. These decomposition methods summarized by Shorrocks and Heshmati have had wide applications among economists who wish to better understand the income inequality issue in the United States.

3.1 By Subgroups

Jenkins (1995), Cowell and Jenkins (1995), and Shorrocks (1983) have developed and popularized an approach to study the factors influencing income inequality by decomposing total income inequality into within and between subgroups' inequality. Not all inequality indices are suitable for this type of decomposition, however. Heshmati (2004) proposed decomposing the Gini index by subgroups. However,

the Gini coefficient does not satisfy the properties of uniform addition (Cowell 1988, 1995; Shorrocks 1980). Furthermore, other widely accepted inequality indices, such as relative mean deviation, the variance and logarithmic variance, also cannot be decomposed by subgroups. However, it can be shown that the Theil index and other Generalized Entropy indices satisfy the uniform addition property and thus admit to a standard decomposition often sought after by researchers (Papatheodorou 2000).

Notwithstanding the interpretation issues due to the failure to adhere to the uniform addition property, Yitzhaki (2002) decomposed the Gini index by the rich and the poor subgroups in Romania in 1993 and calculated income inequalities within and between these subgroups. He found that a poverty index can be derived from a decomposition of an appropriate index of inequality, such that poverty indices are redundant since the decomposition provides additional information to measure poverty. Papatheodorou (2000) used Greek family income data to decompose the Theil and other Generalized Entropy indices into within- and between-subgroups inequality. In his study, he divided the population into subgroups by characteristics of households such as locality, region, age of head, educational level of head, and occupational status of head. He found that except for education, within-group inequality components are larger than between-group inequality components for most population subgroups.

3.2 By Income Sources

The second approach to decompose income inequality focuses the extent to which inequality in different income sources contributes to total income inequality. Shorrock (1983) studied the contribution of various income sources to total income inequality between 1968 and 1977. His research showed that the largest proportion of total income inequality is attributable to labor income and the second largest is attributable to capital earnings. The results from this study, however, differ by the areas studied. For example, El-Osta (1995) found that non-farm income has an equalizing effect on inequality of the United States. For a developing country such as Ecuador, studied by Elbers and Lanjouw (2001), income from non-farm activities increased overall income inequality.

3.3 By Factor Components (Regression-Based Decomposition)

Overcoming many of the limitations to other approaches based on standard decomposition by groups, the decomposition with regression-based approaches is built on decomposition by income sources and has more appealing properties. Through appropriate specification of the explanatory variables, the potential influence on inequality of various factors that otherwise might require separate modeling can

be easily and uniformly incorporated within the same econometric model (Cowell and Fiorio 2011).

To study the cause for cross-country income inequality for OECD countries from 1962 to 1993, Gisbert (2001) decomposed the Theil index by causal factors that include productivity, employment rate and activity rate (reverse causality from income to productivity is acknowledged but not addressed herein). He found from this analysis that while the decrease in productivity inequality across countries causes income inequality to decrease, the employment and activity rate factors have opposite effects on income inequality among OECD countries. Morduch and Sicular (2002) decomposed income inequality by factor components to study the cause of the uneven income distribution in rural China. Morduch and Sicular's results, however, vary substantially based on which indices they use. The decomposition using the Theil index shows that education and demographic variables have strongly reduced inequality. On the other hand, the Gini decomposition indicates that these variables contribute positively, although modestly, to inequality. Baye and Eop (2011) also utilized the regression-based income inequality method to study the cause for income inequality in Cameroon. They found that education is one of the major contributions to income inequality.

Although the results of income inequality decomposition differ by countries and data, these methods provide a useful baseline for studying income inequality in the United States.

4 Data

The data we use to examine the sources of and factors related to income inequality in the U. S. are from Panel Study of Income Dynamics. We use data from the years 1985, 1990, 1995, 1999, 2005, 2009 and 2013. The income of a householder can be roughly categorized into three types: labor income, asset income, and transfers. Labor income includes wages and salaries, bonuses, overtime, tips, commissions, professional practice and trade, market gardening, additional job income, miscellaneous labor income, farm income, and the labor portion of business income. Asset income consists of dividends, interest, rent, trust fund, annuities and the asset portion of business income. Other transfer income is composed of supplemental security income, welfare, pension, and alimony. Other variables that represent the characteristics of the householders are years of education, gender, race, age, marital status, and industries. We drop all the observations indicating non-positive or missing incomes. In addition, we use the selection model to filter out the zero income observations. We use predicted estimators to calculate the inverse mills ratio, which will be used as a new explanatory variable in the income equation. Then, all the zero income observations are dropped.

The number of observations for each year is between 6000 and 8000. Table 1 summarizes information about the head of the household's income for each year. The mean of income increases from $19,931 in 1985 to $45,638 in 2013. Table 2

Table 1 Summary statistics of income of PISD data

	Obs	Mean	Std. Dev	Min	Max
1985	6067	19931.11	26213.84	1	999,999
1990	7897	24110.67	29121.82	5	1,000,000
1995	7536	29710.57	49148.31	1	1,999,992
1999	6317	36095.93	50088.64	1	1,985,000
2005	7208	41749.47	54897.09	1	1,250,000
2009	7779	46038.47	58372.21	1	974,500
2013	7675	45637.98	53685.52	1	990,000

Table 2 Summary statistics of PISD data

	Education	Age	White %	Male %	Married %
1985	12.18	44.62	65.40	74.95	59.40
1990	12.31	45.65	66.43	75.65	58.95
1995	12.72	43.10	62.82	71.66	53.17
1999	12.86	43.55	62.48	72.27	54.82
2005	13.07	43.79	62.47	71.84	52.12
2009	13.38	43.89	61.22	71.02	49.40
2013	13.58	43.82	59.58	70.80	47.91

provides more demographic information for the head of the household. The mean of education years increases from 12.18 in 1985 to 13.58 in 2013. The average age of people in this data decreases from 44.6 to 43.8. In addition, the percentages of white people, the male and married people in this survey all decrease. Table 3 summarizes the share of total population from each industry. There are 13 industries in the present analysis:

Table 3 shows that the manufacturing industry, wholesale and retail trade industry, and professional and related services industry have the largest numbers of employees. In addition, there are a moderate number of people who do not work for money in this population and they are grouped into industry 0. Wee decompose income inequality of this population by subgroups, income sources and factor components with regression.

5 Empirical Methodology

5.1 Decomposition by Subgroups

We now present the indices that are suitable for decomposition by the population subgroup approach (Bourguignon 1979; Cowell 1980, 1988, 1995; Shorrocks 1980; Anand 1983). Generalized entropy indices, denoted as E_θ, come from a family of indices that is decomposable by population subgroup (Shorrocks 1983, Cowell 1995). The generalized entropy indices E_θ can be written as:

Table 3 Percentage of totally population in the industry (ind0-ind12)

	Ind1	Ind2	Ind3	Ind4	Ind5	Ind6	Ind7
1985	2.88	0.87	6.94	20.2	7.57	13.7	3.43
1990	3.19	0.66	7.48	19.3	6.83	13.6	3.91
1995	2.32	0.52	6.67	16.1	6.9	13.6	3.73
1999	2.56	0.44	7.00	16.6	7.46	12.4	4.37
2005	2.29	0.29	8.32	13.4	6.16	12.1	5.08
2009	2.03	0.6	8.45	12.2	6.12	10.7	5.26
2013	2.07	0.66	7.32	12.2	6.12	11.4	5.12

Summary statistics of PISD data (Cont.)

	Ind8	Ind9	Ind 10	Ind 11	Ind 12	Ind 0
1985	4.25	3.4	0.69	12.6	6.63	16.8
1990	4.98	3.19	0.75	12.6	6.42	17.1
1995	4.79	2.64	0.98	15.1	6.81	19.9
1999	6.13	2.79	1.01	15.0	7.08	17.2
2005	8.84	8.42	1.48	14.8	5.95	12.9
2009	10.0	9.35	1.58	14.9	6.45	12.3
2013	10.4	10.8	1.51	15.7	6.63	10.2

Industry 1: Agriculture, Forestry, and Fisheries.
Industry 2: Mining.
Industry 3: Construction.
Industry 4: Manufacturing.
Industry 5: Transportation, Communications, and Other Public Utilities.
Industry 6: Wholesale and Retail trade.
Industry 7: Finance, Insurance, and Real estate.
Industry 8: Business and Repair Services.
Industry 9: Personal Services.
Industry 10: Entertainment and Recreation Services.
Industry 11: Professional and Related Services.
Industry 12: Public Administration.
Industry 0: not working for money now at all.

$$E_\theta = \frac{1}{\theta\,(1-\theta)}\left[\frac{1}{n}\sum\nolimits_i \left(\frac{y_i}{\mu}\right)^\theta - 1\right], \tag{1}$$

where the parameter θ can take any positive, zero or negative value. μ is the mean level of income of the population, and n is the population.

Total income inequality indices in this population subgroup can be decomposed into between- and within-subgroup inequality, which can be written as:

$$I_T = I_B + I_w, \tag{2}$$

where I_T is the overall inequality of the population. I_W is within-group inequality and I_B is between-group inequality.

Between-group inequality can be represented as:

$$I_B = \frac{1}{\theta\,(1-\theta)}\left[\sum\nolimits_k \frac{n_k}{n}\left(\frac{\mu_k}{\mu}\right)^{\theta} - 1\right], \tag{3}$$

while within-subgroup inequality can be calculated by:

$$I_w = \sum\nolimits_K \left(\frac{n_k}{n}\frac{\mu_k}{\mu}\right)^{\theta}\left(\frac{n_k}{n}\right)^{1-\theta} I_k, \tag{4}$$

where n_k is the population in group k, μ_k is the mean income in group k, and I_k is the inequality in group k.

5.2 Decomposition by Income Sources

Stark, Taylor and Yitzhaki (1986) proposed that the influence of any income source on total income inequality depends on how much the income source contributes to income, how equally or unequally distributed the income source is, and how the distribution of the income source and the distribution of total income are correlated. They express the composite Gini coefficient as:

$$G = \sum\nolimits_k S_k G_k R_k, \tag{5}$$

where S_k is the share of source k in total income, G_k is the Gini coefficient corresponding to the distribution of income from source k, and R_k is the correlation of income from source k with the distribution of total income.

5.3 Decomposition by Factor Components (Regression-Based Decomposition)

Using 'natural decomposition rules', inequality indices can be written as a weighted sum of incomes (Shorrocks 1982):

$$I(\mathbf{y}) = \sum a_i(\mathbf{y})\, y_i. \tag{6}$$

The proportional contribution of source k to overall inequality can be written as:

$$s^k = \frac{\sum_{i=1}^{n} a_i(\mathbf{y})\, y_i^k}{I(\mathbf{y})}. \tag{7}$$

The regression-based decomposition approach brings together inequality decomposition by income source and decomposition by population subgroup. We begin with the income equation (Morduch and Sicular, 2002):

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{8}$$

where $\boldsymbol{X}$ is an $n \times M$ matrix of independent explanatory variables, $\boldsymbol{\beta}$ is an M-vector of regression estimates, and $\boldsymbol{\varepsilon}$ is an n-vector of residuals. The coefficient $\boldsymbol{\beta}$ can be estimated using appropriate econometric model and the prediction of $\boldsymbol{y}$ can be formed by $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. In this model, we can estimate the income flows attributed to different household factors by $\widehat{\boldsymbol{y}}^{\boldsymbol{m}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{m}}$, where the household variables include education, age, marriage etc. Therefore, the total income of individual i can be written as the sum of income flows from different source:

$$y_i = \sum\nolimits_{m=1}^{M+1} \widehat{\boldsymbol{y}}_i^m, \text{ for all } i. \tag{9}$$

where $\widehat{\boldsymbol{y}}_i^m = \widehat{\boldsymbol{\beta}}_{\boldsymbol{m}} x_i^m$, for $m = 1, \ldots, M$ and $\widehat{\boldsymbol{y}}_i^m = \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{i}}$, for $m = M + 1$. The share of variable m is then modified to be:

$$s^m = \widehat{\beta}_m \left(\frac{\sum_{i=1}^n a_i(\boldsymbol{y})\, x_i^m}{I(\boldsymbol{y})} \right), \text{for } m = 1, \ldots, M. \tag{10}$$

Since the decompositions above are linear in the estimated parameters, the standard error of s^m can be obtained by:

$$\sigma(s^m) = \sigma\left(\widehat{\beta}_m\right) \left[\frac{\sum_{i=1}^n a_i(\boldsymbol{y})\, x_i^m}{\boldsymbol{I}(\boldsymbol{y})} \right]. \tag{11}$$

6 Results

6.1 Decomposition by Subgroup

Using the decomposition by subgroup method, we categorize the total population by education level, gender and race. Education levels include primary level, secondary level and above secondary level. Genders consist of males and females. White and non-white are used to categorize the total population by races.

Table 4 Decomposition of inequality by education subgroups

	Popn. Share	Mean	Income Share	GE(0)	GE(1)	Gini
By Education						
1985						
Primary	0.026	8762	0.011	0.88	0.52	0.54
Secondary	0.622	18,089	0.565	0.53	0.41	0.46
Above Secondary	0.352	23,998	0.424	0.45	0.38	0.44
Within Groups				0.51	0.40	
Between Groups				0.016	0.015	
2013						
Primary	0.010	21,894	0.005	0.465	0.25	0.371
Secondary	0.400	30,355	0.263	0.639	0.397	0.474
Above Secondary	0.594	56,336	0.732	0.589	0.426	0.482
Within Groups				0.608	0.417	
Between Groups				0.045	0.042	

Differences in equality have been found among the subgroups categorized by educational level as we can see in Table 4. An increasing number of people receive above secondary education from 1985 to 2013. In 1985, people who receive above-secondary education only earn on average about $6000 more than the secondary school graduates. This number jumps to $26,000 in 2013, reflecting the effect of high-level education on the earnings.

The group with the highest within-group income inequality in 1985 contains those with primary school as their highest education attainments. However, in 2013 inequalities within both secondary and above secondary group exceed the inequality within primary school group. One explanation may lie in the fact that these two latter groups contain a large number of individuals who have undergone further training and acquired a variety of certificates, diplomas and degrees. These different levels of training within the groups create greater income disparities (Okatch 2012). Moreover, inequality between these three subgroups almost triples, as it increases from 0.015 to around 0.045. Therefore, we can conclude that educational disparities do increase inequality. However, for all indices between-group inequality is much lower than within-group inequality, so one would need to identify more factors within the subgroups to explain the pervasive inequality we observe in America.

Some researchers and policy-makers have emphasized the income difference between males and females. Table 5 shows results based on the decomposition by gender subgroups from 1985 to 2013. In 1985, women only contribute to 13% of total income, which is much lower than their 25% share of population. By 2013, females' share of total income increases to 29%, which is roughly the same as their population share in the workforce. This result also points to a shrinking, but still significant, gender income gap, which is consistent with decreasing between-group inequality.

Table 5 Decomposition of inequality by gender subgroups

	Popn. Share	Mean	Income Share	GE(0)	GE(1)	Gini
By Gender						
1985						
Male	0.75	23,097	0.87	0.434	0.37	0.43
Female	0.25	10,459	0.13	0.6	0.38	0.47
Within Groups				0.474	0.37	
Between Groups				0.051	0.043	
2013						
Male	0.708	52,525	0.815	0.608	0.43	0.483
Female	0.292	28,939	0.185	0.653	0.429	0.492
Within Groups				0.621	0.43	
Between Groups				0.0335	0.0302	

Table 6 Decomposition of inequality by race subgroups

	Popn. Share	Mean	Income Share	GE(0)	GE(1)	Gini
By Races						
1985						
Non-white	0.346	13,577	0.236	0.49	0.32	0.43
White	0.654	23,292	0.764	0.5	0.40	0.45
Within Groups				0.49	0.38	
Between Groups				0.03	0.03	
2013						
Non-white	0.404	32,480	0.288	0.56	0.38	0.46
White	0.596	54,563	0.712	0.67	0.45	0.49
Within Groups				0.62	0.43	
Between Groups				0.03	0.03	

While inequality across gender groups decreases, inequality within each gender group rises over years of our study. The within-group GE(0) increases from 0.474 to 0.621, and the GE(1) rises from 0.367 in 1985 to 0.43 in 2013. GE(0) that is more sensitive to the population at the bottom of income distribution, increased more than GE(1) (also called T-Theil), which is more sensitive to the population with median income. This reflects greater inequality in the group of people with low income. Moreover, all the inequality measures for 1985 and 2013 suggest that income inequality among women is higher than that among men. One possible reason for this is that some women continue pursuing their career after completing their formal education, while some choose to work in the home. This disparity in career choice leads the greater income gap within the female group.

Table 6 displays results based on the decomposition into white and nonwhite subgroups. Whites always earn more than nonwhites, as they contribute 76.4% of total income in 1985 and 71.2% in 2013, but the percentages decrease. The decrease in the income share of whites is partially explained by the decrease in the population share of whites, which falls from 65.4% to 59.6% between 1985

and 2013. Inequality increases from 0.47 to 0.56 within the nonwhite subgroup and from 0.50 to 0.67 for whites using GE(0). When measured by GE(1), inequality within the nonwhite subgroup increases from 0.32 in 1985 to 0.38 in 2013, and among the white subgroup, this increases from 0.40 to 0.45. The between-group income inequality across those two groups remains constant from 1985 to 2013, but the total within-group inequality values measured by GE(0), GE(1), and Gini all increase. We find that rising inequality within these two race subgroups contributes much more to the increase in overall income inequality than rising inequality across these two racial subgroups.

6.2 Decomposition by Income Sources

Table 7 provides the results of income decomposition by income sources. Recall that total income is categorized into three parts: labor income, asset income and other transfers. From 1985 to 2013 labor income contributes most to total income. Interestingly, labor income is the most evenly distributed income source among these three. With its large share in total income and its relatively even distribution, the labor income has negative numbers under the '% change' column that shows the percentage of increase in inequality caused by certain income source, reflecting that labor income has an equalizing effect on income distribution.

Among these three sources, asset income, which includes dividends, interest, rent, trust fund, annuities and the asset portion of business income, is the only source that constantly positively contributes to income inequality. Its G_k is slightly lower than other transfers but still at a high level. Moreover, asset income's contribution to total income is much larger than that of other transfers. From 1985 to 2013, the portion of labor income increases while asset income and other transfers take fewer shares. The contribution of asset income to total income inequality drops due to the decrease of its portion in total income.

Other transfers have the highest level of inequality, with a G_k of around 0.99. However, the correlation between the distribution of other transfers and the distribution of total income is relatively low, meaning that low-income classes can still have high other transfers. Some people do not work for money, have no capital gains, and thus are at the bottom of income distribution. However, they may receive supplemental security income and well as other transfers. The direction of other transfers' effect on income inequality changes year to year but given its small contribution to total income the magnitude is not significant.

Table 7 Decomposition of inequality by income sources

Share	S_k	G_k	R_k	Share	% Change
1985					
Labor Income	0.8793	0.4701	0.9466	0.8526	−0.0268
Asset Income	0.1165	0.9079	0.6397	0.1474	0.0309
Other Transfer	0.0042	0.9882	0.0031	0.0000	−0.0041
Total Income		0.4590			
1990					
Labor Income	0.8684	0.4721	0.9481	0.8365	−0.0318
Asset Income	0.1275	0.9845	0.6557	0.1627	0.0353
Other Transfer	0.0042	0.9799	0.0844	0.0007	−0.0034
Total Income		0.4646			
1995					
Labor Income	0.9003	0.5165	0.9707	0.8729	−0.0275
Asset Income	0.089	0.9663	0.6845	0.1138	0.0248
Other Transfer	0.0107	0.9916	0.6504	0.0133	0.0026
Total Income		0.5171			
1999					
Labor Income	0.9111	0.4865	0.9644	0.8860	−0.0251
Asset Income	0.0824	0.9541	0.6545	0.1966	0.0242
Other Transfer	0.0065	0.9915	0.5546	0.0074	0.0009
Total Income		0.4825			
2005					
Labor Income	0.9196	0.5101	0.9762	0.9241	−0.0085
Asset Income	0.0768	0.9499	0.6274	0.0758	0.0119
Other Transfer	0.0036	0.9828	0.0178	0.0001	−0.0034
Total Income		0.5026			
2009					
Labor Income	0.9326	0.5101	0.9762	0.9241	−0.0085
Asset Income	0.0639	0.9499	0.6274	0.0758	0.0119
Other Transfer	0.0035	0.9828	0.0178	0.0001	−0.0034
Total Income		0.5026			
2013					
Labor Income	0.9345	0.5115	0.9726	0.9286	−0.0059
Asset Income	0.0603	0.9568	0.5944	0.0685	0.0082
Other Transfer	0.0052	0.9891	0.2788	0.0029	−0.0024
Total Income		0.5007			

6.3 Decomposition by Factor Component (Regression-Based Decomposition)

Table 8 shows the result of the decomposition of the Gini coefficient by factor components using he regression-based approach for years 1985, 1990, 1995, 1999, 2005,

Table 8 Results of regression-based approach

	1985	1990	1995	1999	2005	2009	2013
Education	21.14	40.87	32.79	21.6	21.3	27.61	20.09
White	11.82	10.06	12.92	7.83	7.2	4.61	5.62
Married	5.43	7.19	4.23	7.78	5.5	8.83	8.08
Male	14.75	12.54	12.44	5.68	7.78	2.87	4.64
Age	−15.72	−11.85	−15.17	−14.8	−6.9	−1.02	−1.04
Age^2	16.01	12.49	18.24	25.29	6.65	0.18	0.33
Industry 1	−6.5	−3.41	0.24	−0.64	−0.53	−0.21	−0.16
Industry 2	3.46	1.60	1.91	0.97	2.29	4.8	3.11
Industry 3	3.21	6.26	7.64	3.95	15.85	13.1	4.35
Industry 4	42.84	30.38	30.82	17.64	35.4	32.41	21.75
Industry 5	22.42	19.14	17.01	11.98	19.03	17.17	8.31
Industry 6	−1.28	1.68	2.38	−2.51	−2.2	−6.88	−8.01
Industry 7	7.55	9.3	7.51	7.02	17.91	20.23	13.56
Industry 8	−0.2	1.36	2.77	3.44	−27.83	−36.51	−22.8
Industry 9	−13.55	−9.39	−6.72	−5.00	−4.99	−6.31	−4.22
Industry 10	0.09	−0.19	1.04	−0.095	−2.41	−2.15	−1.33
Industry 11	9.95	12.3	15.56	5.34	18.71	26.79	10.05
Industry 12	13.19	15.29	15.42	9.64	27.16	31.49	19.58
White male	−14.44	−7.78	−9.2	−2.51	−1.17	3.72	−1.05
Inverse Mills	−44.35	−78.04	−103.02	−57.1	−106.21	−97.77	−41.88
Residual	75.82	30.2	51.19	54.495	67.46	57.04	61.02

2009 and 2013. The detailed OLS regression results and income flow shares are attached in the Appendix. All the coefficients from OLS regression are significant at conventional levels except the coefficients on white males. They are significant at only the 10% level in 1985 and 2013, insignificantly different from zero at the 10% level in 1999 and 2005, and significant at the 1% level in 1990, 1995, and 2009.

From Table 8, we can see that the combined effects of age and age squared is very close to 0 from year 1985 to 2013, except the year 1999 when the age and age square contributes 10.49% to total inequality. However, for most of the years, age is not the main factor that affects income inequality. Education has large positive shares for all the years. Its effect rises to over 30% in 1990s, then decreases to 20%, and stays at a pretty stable level after 1999. This result reflects that the educational gap explains over 20% of income inequality. The dummy variable WHITE provides 11.82% contribution to overall inequality in 1985. Nevertheless, this value decreases over years, and drops to 5.62% in 2013, which shows that even though race still positively affects inequality, its influence has decreased. This result is consistent with the finding by subgroup decomposition that inequality between different race groups explains total income inequality to a smaller extent. The share taken by MARRIED is positive for all years with a slight increase. This is in line with the assortative matching theory in Sect. 2.7 that marriage causes social segregation and income inequality. MALE also explains part of inequality, but its contribution decreases from 14.75% in 1985 to 4.64% in 2013, indicating that gender inequality is not a big cause for current inequality.

Industry 4, the manufacturing industry, provides the most contribution to total inequality among all industries. The manufacturing industry takes a 42.84% share of inequality in 1985, which drops to 17.64% in 1999, and increases again after the 2000s. One of possible explanations for why the manufacturing industry contributes so much to inequality in the early years is that it has the largest number of employees, and thus has a big effect on total income. Moreover, as shown in the tables in Appendix, there are many more employees in this industry clustering around the top 25% class of income distribution than in the bottom 25%. According to Charles Kenny in his Bloomberg article on April 28, 2014, however, the manufacturing industry has shrunk after the 1990s, due in part to the growth in imports from China. Given fewer employees and thus smaller contributions to total income, industry 4 has a lower effect on inequality in the 1990s. One reason for the increase in the share of the manufacturing industry on inequality in the later years might be the displacement of technology. Gary Becker has pointed out in *The Becker-Posner Blog* posted on April 22, 2012 that labor costs will be a lower fraction of the total cost of manufactured products with the advent of new technologies. These technologies are unlikely to create job opportunities since they are generally labor-saving, instead of labor-using. Jobs with new technologies will require skilled and better-paid workers, which magnify income inequality in this industry.

Professional and related service (Industry 11) is another industry that has a large work force and a positive contribution to total income inequality. Industry 11 generally maintains a more-than-10% contribution to income inequality, reaching a peak at 26.79% in 2009. Jobs in this industry include lawyers, engineers, accountants, consultants, doctors and other highly paid professionals. These jobs require high education levels and advanced cognitive or interpersonal skills. They tend to receive relatively high income and may be less likely to be replaced by technology during the sample period.

The contribution provided by industry 7, which is finance, insurance, and real estate industry, rises from 7.55% in 1985 to 20.23% in 2009. The increased impact of this industry is due to the explosive growth of the financial sector. As Sabadish and Mishel have argued, someone who is either an executive or is employed in the financial sector is often in the top 1% income households in the United States. With the large increased incomes in the financial industry and the high correlation between the income distribution of the total population and the income distribution within the financial sector, industry 7 explains a large proportion of total income inequality.

7 Conclusion

The main goal of this paper is to study the causes of income inequality in the United States. In order to quantify the effects of different factors on inequality, we use three types of decomposition approaches: by subgroups, by income sources, and by factor components based on regression.

Using decomposition by subgroups, we find that inequality within the high-level education group increases from 1985 to 2013, indicating that the availability of more professional options for people with above secondary education degrees has resulted in an increased dispersion of income levels within the group. Although contributing much less to total inequality than within-group inequality, inequality between different education groups has expanded. Between-group inequalities along gender and racial lines decrease from 1985 to 2013, possibly reflecting to a certain degree the effectiveness of measures aimed at addressing racial and gender income inequality. Similar to the results of decomposition by education level, the effect of within-group inequality on total income inequality outweighs that of between-group inequality. This suggests that merely focusing on devising policies to address the income gap across different groups may be misguided.

Decomposing inequality by income sources, we find that labor income is more evenly distributed than asset income and other transfers. Asset income has uneven distribution and is highly correlated with the distribution of total income. It positively affects overall income inequality during our sample period.

In the regression-based decomposition analysis, education, marriage, race and gender increase income inequality, but the effects of race and gender have decreased between 1985 and 2013, reflecting in part more equalized income distributions among races and genders. Different industries have influenced income inequality in different directions and to different degrees. The manufacturing industry, finance, insurance and real estate industry, public utilities industry, professional and related services industry, and public administration industry all positively contribute to total inequality.

For all the years of in our sample we find that a substantial percentage of income inequality cannot be explained by the observable factors one typically argues are at the heart of income inequality. The relatively high contribution of the residual's value is consistent with findings from other studies, such as those of Morduch and Sicular (2002), Wan and Zhou (2005) and Yun (2006). For future research, other unobserved factors that may explain total income inequality are worth exploring.

Appendix

Regression Results and Descriptive Statistics For 1985, 1990, 1995, 1999, 2005, 2009, 2013

In the following tables, income shares are calculated by multiplying the mean value of each explanatory variable by its estimated coefficient from the earnings equation. For each quartile in the distribution, shares equal the sum of estimated income flow from each variable over all households in the quartile, divided by the sum of flows from variable m for the entire sample.

Table A.1 Year 1985

	Linear Regression Equation			Shares of Income Flows by Quartile				
	Est coef	Std err	Income Share	Bottom	Second	Third	Top	Ratio
Education	3064.24	218.21	186.18	22.09	23.83	25.78	28.29	1.28
White	15370.17	2267.75	50.44	20.34	21.32	26.16	32.18	1.58
Married	4786.16	956.94	14.26	15.15	21.61	28.27	34.96	2.31
Male	15028.43	2086.29	56.51	17.79	23.20	27.56	31.45	1.77
Age	678.3	78.49	151.84	29.11	24.27	23.04	23.58	0.81
Age^2	−5.33	0.70	−63.26	34	24.28	21.04	20.68	0.61
Industry1	34149.48	5513.22	4.94	40.57	36.57	17.71	5.12	0.17
Industry2	52470.98	6238.58	2.30	0	16.98	30.19	52.83	N/A
Industry3	47092.92	5667.16	16.40	16.15	30.40	31.35	22.09	1.37
Industry4	49986.63	5596.01	50.76	8.31	22.88	33.88	34.93	4.20
Industry5	51174.01	5586.99	19.42	7.41	20.48	25.71	46.41	6.26
Industry6	45932.40	5548.09	31.60	24.28	27.64	27.04	21.03	0.87
Industry7	54404.48	5641.43	9.36	9.13	27.88	23.08	39.9	4.37
Industry8	46602.54	5608.6	9.35	23.26	32.56	22.09	22.09	0.95
Industry9	45332.52	6198.61	7.72	61.65	21.84	11.17	5.33	0.09
Industry10	44658.87	6634.92	1.55	21.43	30.95	28.57	19.05	0.89
Industry11	46602.54	5366.98	29.41	16.25	29.49	27.52	26.74	1.65
Industry12	46845.48	5568.13	15.57	7.46	23.38	36.07	33.08	4.43
White_male	−6377.51	2184.69	−16.76	15.2	19.35	27.53	37.92	2.49
Inv_mills	223098.4	31802.8	364.30	20.2	24.64	23.38	14.16	0.70
Constant	−168239.6	18905.8	−841.90					
R-square	0.1514							

Table A.2 Year 1990

	Linear Regression Equation			Shares of Income Flows by Quartile				
	Est Coef	Std Err	Income Share	Bottom	Second	Third	Top	Ratio
Education	5314.38	217.11	271.38	21.42	23.12	25.8	29.67	1.39
White	18532.97	1722.27	51.06	21.27	21.92	25.52	31.28	1.47
Married	8420.21	933.81	20.59	15.51	22.62	27.73	34.14	2.20
Male	16063.86	1690.38	50.4	17.55	23.72	27.27	31.46	1.79
Age	648.76	127.441	124.72	28.64	24.72	23.02	23.62	0.82
Age^2	−5.22	1.27	−53.09	33.15	25.12	20.85	20.88	0.63
Industry1	97247.61	6540.81	12.87	32.94	30.16	17.06	19.84	0.60
Industry2	97721.43	7156.99	2.67	9.62	13.46	36.54	40.38	4.20
Industry3	100560.2	6375.42	31.21	15.4	28.6	29.27	26.73	1.74
Industry4	99427.55	6285.51	79.53	10.05	26.66	32.63	30.66	3.05
Industry5	100106.7	6235.72	28.34	7.61	19.29	28.01	45.08	5.92
Industry6	95028.59	6232.58	53.45	22.13	29.97	24.56	23.34	1.05
Industry7	98688.56	6225.88	16.02	8.09	24.92	25.89	41.1	5.08
Industry8	94078.49	6199.06	19.42	20.61	27.48	27.73	24.17	1.17
Industry9	94287.38	6678.33	12.5	50.79	30.16	11.51	7.54	0.15
Industry10	98138.85	7287.10	3.04	18.64	42.37	20.34	18.64	1.00
Industry11	92924.55	5987.46	48.61	14.56	28.01	29.92	27.51	1.89
Industry12	94193.10	6178.35	25.08	5.92	17.95	40.04	36.09	6.10
White_male	−8939.70	1820.70	−19.1	15.51	20.91	26.91	36.67	2.36
Inv_mills	456553.1	33323.1	617.57	29.36	24.66	23.32	14.34	0.49
Constant	−314705.6	18757.3						
R-square	0.2283							

Table A.3 Year 1995

	Linear Regression Equation			Shares of Income Flows by Quartile				
	Est Coef	Std Err	Income Share	Bottom	Second	Third	Top	Ratio
Education	7066.66	425.31	305.61	22.01	23.68	25.78	28.53	1.30
White	33268.94	3722.24	70.34	21.14	21.17	26.34	31.35	1.48
Married	6145.94	1518.75	11	14.75	19.27	29.95	36.04	2.44
Male	21554.59	2892.22	51.99	24.73	17.87	28.56	31.44	1.27
Age	1314.944	204.00	190.78	29.91	22.01	23.35	24.73	0.83
Age^2	−11.93	2.08	−83.81	36.76	19.69	20.58	22.97	0.62
Industry1	113462.6	11348.08	8.87	22.29	30.29	21.14	26.29	1.18
Industry2	134565.8	14114.14	2.34	2.56	15.38	35.9	46.15	18.03
Industry3	123370.8	11740.82	27.72	11.73	26.44	34.19	27.63	2.36
Industry4	129076.2	11922.23	69.93	6.27	25.89	35.2	32.65	5.21
Industry5	129774.9	12086.40	30.14	7.88	19.62	31.15	41.35	5.25
Industry6	120969.8	11541.55	55.22	17.12	33.17	27.1	22.6	1.32
Industry7	127822.9	11818.79	16.04	9.25	27.76	26.33	36.65	3.96
Industry8	118944.5	11284.56	19.18	18.01	32.13	19.94	29.92	1.66
Industry9	120668.7	12547.71	10.72	44.22	32.66	15.58	7.54	0.17
Industry10	131447.2	12807.75	4.34	12.16	28.38	29.73	29.73	2.44
Industry11	124966.3	11539.37	63.4	13.73	28.87	29.31	28.08	2.05
Industry12	125065.7	11962.51	28.66	7.02	17.93	35.28	39.77	5.67
White_male	−14894.25	3227.03	−24.79	15.78	19.16	28.49	36.57	2.32
Inv_mills	834787.6	88164.66	913.61	29.63	23.96	23.64	14.97	0.51
Constant	−496781.4	44977.81						
R-square	0.1332							

Table A.4 Year 1999

	Linear Regression Equation			Shares of Income Flows by Quartile				
	Est Coef	Std Err	Income Share	Bottom	Second	Third	Top	Ratio
Education	5331.55	381.3942	189.96	22.68	23.14	25.42	28.75	1.27
White	22817.15	4537.433	39.5	21.89	21.61	23.99	32.51	1.49
Married	13745.05	1803.548	20.88	15.56	20.79	27.78	35.86	2.30
Male	11343.24	2970.838	22.71	18.27	22.56	27.6	31.57	1.73
Age	2847.657	264.9826	343.56	28.44	21.62	24.38	25.55	0.90
Age^2	−32.26112	3.304188	−227.81	27.72	15.85	35.94	20.49	0.74
Industry1	89542.59	14019.1	6.36	24.69	32.72	22.84	19.75	0.80
Industry2	100267.7	16931.77	1.23	7.14	10.71	35.71	46.43	6.50
Industry3	89598.28	13824.94	17.37	14.93	28.51	26.92	29.64	1.99
Industry4	94641.96	14323.07	43.42	9.56	26.96	31.07	32.41	3.39
Industry5	95166.51	14059.42	19.66	8.49	19.53	30.57	41.4	4.88
Industry6	86675.76	14151.57	29.65	23.08	32.95	25.13	18.85	0.82
Industry7	100365.9	14322.72	12.15	9.78	21.38	28.62	40.22	4.11
Industry8	98006.31	13459.75	16.63	20.41	23.77	25.58	30.23	1.48
Industry9	87022.04	15379.37	6.72	51.17	25.57	18.75	4.55	0.09
Industry10	88962.71	15425.79	2.5	21.88	29.69	31.25	17.19	0.79
Industry11	89177.06	14319.17	37.15	17.79	29.79	27.05	25.37	1.43
Industry12	88751.8	14305.77	17.4	6.04	21.7	35.57	36.69	6.07
White_male	−4593.461	3670.839	−6.27	16.35	19.85	25.86	37.94	2.32
Inv_mills	685027.4	128709.2	604.22	29.04	23.53	24.11	15.06	0.52
Constant	−410793.7	60796.74						
R-square	0.1649							

Table A.5 Year 2005

	Linear Regression Equation			Shares of Income Flows by Quartile				
	Est Coef	Std Err	Income Share	Bottom	Second	Third	Top	Ratio
Education	10004.84	463.8524	209.27	23.42	23.91	25.59	27.09	1.16
White	27161.06	2906.205	19.39	21.81	22.12	24.74	31.33	1.44
Married	11517.64	1628.997	12.70	15.01	19.75	29.28	35.96	2.40
Male	20135.41	2479.82	16.31	18.44	22.98	27.6	30.98	1.68
Age	2084.26	220.3754	161.13	27.9	22.52	24.03	25.55	0.92
Age^2	−18.66558	2.304441	−54.88	30.84	19.17	26.52	23.48	0.76
Industry1	301,587	21790.82	10.75	29.7	27.27	18.79	24.24	0.82
Industry2	305138.9	24211.51	1.40	0	9.52	23.81	66.67	N/A
Industry3	293155.2	21595.65	37.62	12.83	27.67	30.33	29.17	2.27
Industry4	297290.3	21802.99	61.61	10.35	26.6	31.26	31.78	3.07
Industry5	300009.3	22071.49	28.43	10.81	22.3	32.88	34	3.15
Industry6	290764.5	21736.05	53.53	22.85	28.59	29.05	19.52	0.85
Industry7	304466.9	21546.24	24.51	10.66	23.77	27.87	37.7	3.54
Industry8	287167.2	21726.53	17.23	27.56	33.57	25.09	13.78	0.50
Industry9	303,021	21784.13	84.21	26.53	32.34	22.53	18.6	0.70
Industry10	281632.5	21850.84	6.33	30.84	25.23	29.91	14.02	0.45
Industry11	282345.5	21200.55	50.51	18.91	21.77	24.38	34.95	1.85
Industry12	288763.8	21492.91	26.65	4.43	20.05	33.57	41.96	9.47
White_male	−2821.499	2909.776	6.13	17.06	20.36	26.63	35.95	2.11
Inv_mills	1,564,842	122154.7	581.04	29.06	24.12	23.74	14.52	0.50
Constant	−928940.4	63552.31						
R-square	0.1568							

Table A.6 Year 2009

	Linear Regression Equation			Shares of Income Flows by Quartile				
	Est Coef	Std Err	Income Share	Bottom	Second	Third	Top	Ratio
Education	10765.94	438.79	312.79	22.97	23.36	25.85	27.82	1.21
White	21046.96	1631.60	24.28	21.57	21.19	25.49	31.75	1.47
Married	18257.11	2535.15	22.58	15.64	18.92	28.36	37.08	2.37
Male	9016.56	2187.51	13.91	19.51	22.48	27.04	30.97	1.59
Age	376.21	48.89	35.87	27.61	22.39	24.3	25.71	0.93
Age^2	−0.38	0.08	−1.9	30.18	19.43	26.54	26.85	0.79
Industry1	330934.9	22013.84	14.6	21.52	36.71	20.25	21.52	1.00
Industry2	377422.9	22994.77	4.95	2.13	12.77	27.66	57.45	26.97
Industry3	328990.6	21720.66	60.35	15.68	28.31	27.4	28.61	1.82
Industry4	327030.7	21519.31	86.48	10.67	24.5	33.16	31.68	2.97
Industry5	328333.2	21713.56	43.64	11.76	20.38	36.76	31.09	2.64
Industry6	321773.5	21684.89	75.02	22.04	35.33	23.11	19.52	0.89
Industry7	336969.6	21142.37	38.48	10.51	21.52	27.87	40.1	3.82
Industry8	324591.1	22094.86	70.69	39.10	33.08	18.97	8.85	0.33
Industry9	326452.6	21248.5	66.27	25.03	31.22	23.93	19.81	0.67
Industry10	303233.1	21485.7	10.41	27.64	31.71	25.2	15.45	0.56
Industry11	313692.2	20843.13	101.78	16.61	23.06	29.43	30.90	2.13
Industry12	321851.2	21207.37	45.11	7.37	13.15	33.47	46.02	6.24
White_male	10303.47	2712.92	10.92	17.66	19.35	26.83	36.83	1.79
Inv_mills	1,557,035	112926.3	1075	28.66	23.79	24.07	15.30	0.51
Constant	−926876.7	60985.87						
R-square	0.2166							

Table A.7 Year 2013

	Linear Regression Equation			Shares of Income Flows by Quartile				
	Est Coef	Std Err	Income Share	Bottom	Second	Third	Top	Ratio
Education	8925.10	361.74	265.63	24.08	23.02	25.49	27.40	1.14
White	20833.19	2359.80	27.20	21.78	20.29	25.28	32.65	1.50
Married	19421.08	1488.70	20.39	16.10	18.17	28.37	37.37	2.32
Male	15893.45	2168.87	24.66	20.70	21.97	26.76	30.57	1.48
Age	1188.69	244.64	114.12	27.50	22.27	23.95	26.29	0.96
Age^2	−8.27	2.66	−41.32	29.32	19.37	21.12	30.19	1.03
Industry1	199510.4	16706.5	9.06	22.01	30.82	28.93	18.24	0.83
Industry2	217222.6	17328.21	3.16	1.96	7.84	31.37	58.82	30.01
Industry3	196119.9	16607.4	31.47	18.15	26.51	28.50	26.87	1.48
Industry4	205820.2	16645.64	54.82	12.00	23.15	32.37	32.48	2.71
Industry5	203011.8	16807.62	27.24	14.89	22.77	32.55	29.79	2.00
Industry6	192523.9	16616.45	48.15	27.74	32.31	21.92	18.04	0.65
Industry7	213701.5	16413.12	23.98	10.94	20.87	27.48	40.71	3.72
Industry8	195163.4	16911.02	44.24	40.05	31.99	17.88	10.08	0.25
Industry9	203832.6	16763.34	48.30	24.46	32.05	24.10	19.40	0.79
Industry10	177873.6	16749.65	5.89	30.17	27.59	29.31	12.93	0.43
Industry11	191041.3	16275.06	65.61	22.53	21.78	25.60	30.09	1.34
Industry12	200,469	16542.88	29.13	7.07	14.54	34.77	43.61	6.17
White_male	−2901.55	2571.52	2.99	17.97	18.74	26.09	37.21	2.07
Inv_mills	654915.5	61255.92	464.85	29.26	23.41	23.95	16.12	0.55
Constant	−532260.3	37439.95						
R-square	0.2475							

References

Anand, S. (1983). *Inequality and poverty in Malaysia: Measurement and decomposition*. New York: Oxford University Press.

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives, 29*(3), 3–30.

Autor, D. H., Katz, L. F., & Kearney, M. S. (2008). Trends in U.S. wage inequality: Revising the revisionists. *Review of Economics and Statistics, 90*, 300–323.

Barth, E., Bryson, A., Davis, J. C., & Freeman, R. B. (2016). It's where you work: Increases in earnings dispersion across establishments and individuals in the U.S. *Journal of Labor Economics, 34*(S2), 67–97.

Baye, F. M., Epo, B. N. (2011). *Inequality decomposition by regressed-income sources in Cameroon*. Paper presented at the annual meeting for the Special IARIW-SSA Conference on Measuring National Income, Wealth, Poverty, and Inequality in African Countries, Cape Town, South Africa, September 28–October 1.

Bourguignon, F. (1979). Decomposable income inequality measures. *Econometrica, 47*(4), 901–920. https://doi.org/10.2307/1914138.

Brynjolfsson, E., & McAfee, A. (2011). *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*. Lexinton: Digital Frontier Press.

Card, D., Heining, J., & Kline, P. (2013). Workplace heterogeneity and the rise of west German wage inequality. *The Quarterly Journal of Economics, 128*(3), 967–1015.

Cardarellii, R., & Lusinyan, L. (2015). *U.S. Total factor productivity slowdown: Evidence from the U.S. States*. Washington, DC: International Monetary Fund.

Cortes, G. M., Jaimovich, N., Nekarda, C. J., & Siu, H. E. (2014). The micro and macro of disappearing routine jobs: A flows approach. NBER Working Paper, No. 20307.

Council of Economic Advisors. (2015). *Gender pay gap: Recent trends and explanations*. Council of Economic Advisers issue brief: 1–8.

Cowell, F. A. (1980). On the structure of additive inequality measures. *The Review of Economic Studies, 47*(3), 521–531. https://doi.org/10.2307/2297303.

Cowell, F. A. (1988). Inequality decomposition: Three bad measures. *Bulletin of Economic Research, 40*(4), 309–312.

Cowell, F. A. (1995). *Measuring inequality*. London: Prentice Hall/Harvester Wheatsheaf.

Cowell, F. A., & Fiorio, C. V. V. (2011). Inequality decompositions - a reconciliation. *Journal of Economic Inequality, 9*(4), 509–528. https://doi.org/10.1007/s10888-011-9176-1.

Cowell, F. A., & Jenkins, S. P. (1995). How much inequality can we explain? A methodology and an application to the United States. *The Economic Journal, 105*(429), 421–430. https://doi.org/10.2307/2235501.

Cowen, T. (2013). *Average is over: Powering America beyond the age of the great stagnation*. New York: Dutton Adult.

DeNavas-Walt, C., & Proctor, B. D. (2014). *Income and poverty in the United States: 2013*. Washington, DC: U.S. Government Printing Office.

Eika, L., Mogstad, M., & Zafar, B. (2014). *Educational assortative mating and household income inequality (No. w20271)*. National Bureau of Economic Research.

Elbers, C., & Lanjouw, P. (2001). Inter-sectoral transfer, growth, and inequality in rural Ecuador. *World Development, 29*(3), 481–496.

El-Osta, H. S., Andrew Bernat, G., Jr., & Ahearn, M. C. (1995). Regional differences in the contribution of off-farm work to income inequality. *Agricultural and Resource Economics Review, 24*, 1–14.

Gisbert, F. J. G. (2001). On factor decomposition of cross-country income inequality: Some extensions and qualifications. *Economics Letters, 70*(3), 303–309.

Goldin, C., & Katz, L. F. (2008). *The race between education and technology*. Cambridge, MA: Belknap Press of Harvard University Press.

Heshmati, A. (2004). *Global trends in income inequality*. Hauppauge, NY: Nova Science Publishers.

Jenkins, S. (1995). Accounting for inequality trends: Decomposition analyses for the UK, 1971-86. *Economica, 62*(245), 29–63. https://doi.org/10.2307/2554775.

Kalmijn, M. (1991). Shifting boundaries: trends in religious and educational homogamy. *American Sociological Review, 56*, 786–800.

Kearney, M. S., Hershbein, B., & Boddy, D.. (2015). The future of work in the age of the machine. A Hamilton Project Framing Paper.

Mishel, L. R., Bivens, J., Gould, E., & Shierholz, H. (2012). *The state of working America*. New York: ILR Press.

Morduch, J., & Sicular, T. (2002). Rethinking inequality decomposition, with evidence from rural China. *The Economic Journal, 112*(476), 93–106. https://doi.org/10.1111/1468-0297.0j674.

Okatch, Z. (2012). *Determinants of income inequality in Botswana: A regression-based decomposition approach*. Master's thesis, University of Western Australia.

Rosenfeld, M. J. (2008). Racial, educational and religious endogamy in the United States: a comparative historical perspective. *Social Forces, 87*(1), 1–31.

Papatheodorou, C. (2000). *Decomposing inequality in Greece: Results and policy implications (No. 49)*. LSE STICERD.

Sabadish, N., & Mishel, L. (2012). *CEO pay and the top 1%: How executive compensation and financial-sector pay have fueled income inequality (No. 331)*. Economic Policy Institute.

Schwartz, C. R. (2010). Earnings inequality and the changing association between spouses' earnings. *American Journal of Sociology, 115*(5), 1524–1557.

Schwartz, C. R., & Mare, R. D. (2005). Trends in educational assortative marriage from 1940 to 2003. *Demography, 42*(4), 621–646.

Shorrocks, A. F. (1980). The class of additively decomposable inequality measures. *Econometrica, 48*(3), 613–625.

Shorrocks, A. F. (1982). Inequality decomposition by factor components. *Econometrica, 50*(1), 193–211. https://doi.org/10.2307/1912537.

Shorrocks, A. F. (1983). Inequality decomposition by population subgroups. *Econometrica, 52*(6), 1369–1385.

Stark, O., Edward Taylor, J., & Yitzhaki, S. (1986). Remittances and inequality. *The Economic Journal, 96*(383), 722–740. https://doi.org/10.2307/2232987.

Stiroh, K. J. (2002). Information technology and the U.S. productivity revival: What do the industry data say? *American Economic Review, 92*(5), 1559–1576.

Vardi, M. (2016). Smart robots and their impact on employment. Presentation, AAAS Annual Meeting, Washington, DC.

Wan, G., & Zhou, Z. (2005). Income inequality in rural China: Regression-based decomposition using household data. *Review of Development Economics, 9*(1), 107–120.

Wright, E. O. (1978). Race, class, and income inequality. *American Journal of Sociology, 83*(6), 1368–1397.

Yitzhaki, S. (2002). Do we need a separate poverty measurement? *European Journal of Political Economy, 18*(1), 61–85.

Yun, M.-S. (2006). Earnings inequality in the USA, 1969-99: Comparing inequality using earnings equations. *Review of Income and Wealth, 52*(1), 127–144.

Frontier Estimation of a Cost Function System Model with Local Least Squares: An Application to Dutch Secondary Education

Jos L. T. Blank

Abstract In this paper, we propose a method for cost efficiency measurement that is based on local estimation in several stages. The method is based on weighted least squares where weights depend on the distance of an observation to all other observations and on the distance to the cost frontier. The new element in the method is that it also includes the information from the cost share equations and includes cost efficiency in the weighting matrix. The latter is derived from a first stage and implemented in a second stage analysis. An application to a data set of Dutch school boards in secondary education shows that it works well in practice. It produces a number of reliable estimates. It also shows a variation in outcomes that would be hard to cover with, for instance, traditional procedures such as SFA on a translog cost function.

Keywords Local estimation · Cost efficiency · Scale economies · Technical change · Education

JEL Classification: C01, D24, I21

1 Introduction

Stochastic frontier analysis (SFA) and data envelopment analysis (DEA) are very popular methods for establishing the (cost) efficiency scores of firms. Both methods have been extensively applied to firms in various industries in order to get an insight into the relative (cost) efficiency of individual firms. The methods have also been applied in order to compare the performance of departments within firms, and even to compare the performance of different countries.

J.L.T. Blank (✉)
Delft University of Technology, Delft, The Netherlands

Erasmus University Rotterdam, Rotterdam, The Netherlands
e-mail: j.blank@ipsestudies.nl

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_4

SFA, which was developed by Aigner et al. (1977) and Meeusen & Van den Broeck (1977), is a parametric method. The standard cost or production function is estimated by maximum likelihood methods where the error component consists of random noise and a random (cost) efficiency component, which can be separated empirically. Extensive reviews of the SFA approach can be found in Fried et al. (2008), Kumbhakar & Lovell (2000), Coelli et al. (2005), Blank (2000), and Parmeter and Kumbhakar (2014).

DEA is a technique based on linear programming. This technique is derived from early production work by Farrell (1957) and Debreu (1951) and was later formalised using linear programming techniques (R. D. Banker et al. 1984; Charnes et al. 1978; Färe et al. 1986). The objective of this approach is to envelop the data points as closely as possible, and to produce the best practice frontier by linking several line segments together. This technique thus identifies the efficient observations and calculates the (cost) efficiency scores by measuring the distance to these efficient observations or convex combinations of them.

Both methods have pros and cons. For several decades, each method has been seriously criticised by proponents of the other. The critics of SFA focus on the required functional specification of the model and the distributional assumptions about the (cost) efficiency component. The critics of DEA focus on the absence of a stochastic component and the difficulty of accounting for environmental variables and deriving economic features such as economies of scale and scope and input (or output) substitution.

It is generally recognised that the strong point of SFA is that it takes randomness (measurement and specification errors) into account, whereas the strong point of DEA is the flexibility of the production technology, which does not require a general functional specification. DEA is an observation by an observation technique that provides a local estimator.

Only in recent years has there been a tendency in the literature to try to combine the best of both worlds. Kuosmanen (2008) developed a technique that converts a DEA formulation into a stochastic formulation that can be estimated by maximum likelihood techniques. Another approach was developed by Fan et al. (1996), who used standard kernel methods based on maximum likelihood. He applied the stochastic frontier model without the rigidity of a parametric representation of the technology. For an extensive discussion, see Johnson and Kuosmanen (2015) in Ray et al. (2015).

Less criticism is voiced about the fact that SFA has hardly been applied to the full system of equations that can be derived from duality theory. Complicated solutions have been provided by Kumbhakar & Tsionas (2005), based on Bayesian techniques or through the reformulation of the model based on shadow pricing (Blank and Eggink 2004; Kumbhakar 1997; Maietta 2002). Almost all empirical applications of SFA are therefore limited to single equation models.

In this paper, we will present a method that is based on the idea of local estimation and includes the information from the cost share equations. A possible answer to the aforementioned issues is to apply weighted least squares where weights depend on the distance of an observation to all other observations (more or less lookalikes)

and on the distance to the cost frontier. The latter is derived from a first stage and implemented in a second stage analysis. A similar method for deriving cost efficiency scores in the case of a global estimation of a cost function was proposed earlier by Blank & Meesters (2012). To show how the procedure works in practice, it is applied to a data set of Dutch school boards in secondary education.

The outline is as follows. Section 2 describes the underlying model and the estimation procedure for estimating the model. In Sect. 3 the data are described, and in Sect. 4 the results are presented and discussed. Section 5 concludes the paper.

2 Methodology

The methodology can be applied to a production function, a cost function or any other representation of the production technology and some behavioural assumptions. In this paper, we apply a cost function approach. A cost function describes the relationship between minimum costs on one hand and services delivered and resource prices on the other hand. Minimum costs refer to the assumption that a firm is minimising its costs by allocating its resources optimally and using them in the most technically efficient way. The services produced and resource prices are given. Mathematically this can be described by:

$$C = c\,(y, w) = \min_x \{w \cdot x |\, (y, x) \in T\,(x, y)\} \tag{1}$$

With:

C = (minimum) costs;
y = vector services produced;
w = vector of resource prices;
x = vector of resources;
$T(x, y)$ = set of feasible combinations of services produced and resources.

It is known from duality theory that the optimum allocation of resources can be derived from the cost function according to Shephard's Lemma (Shephard 1953):

$$x = \frac{\nabla C}{\nabla w} \tag{2}$$

Equation (2) tells us that the optimal allocation of resource x equals the gradient of the cost function with respect to the corresponding resource prices.

A functional specification is required in order to conduct empirical research. We assume that a mathematical specification $c(y, w, \theta)$ is at hand where costs are dependent on the aforementioned services produced, resource prices and a set of parameters θ. If data are available, then the parameter set θ can be established by an estimating technique. However, in practice, no data are available on minimum cost. Instead, observable costs may differ from minimum costs due to the (common)

measurement error and to managerial errors. Whereas the first type of errors may go either way, negative or positive, the second type only shows positive values. Mismanagement can only lead to higher costs than the minimum cost. For reasons of convenience we transform all the data to logarithms. In terms of the observable cost we may then rewrite (1) as:

$$\ln(C) = c\,(\ln(y), \ln(w), \theta) + u + v \tag{3}$$

With:

u = (percentage) extra cost due to mismanagement (cost inefficiency);
v = measurement error.

The corresponding share equations are (note that Eq. (2) is now expressed in terms of cost shares due to the logarithmic transformation):

$$S = \frac{\nabla \mathrm{c}\,(\mathrm{y}, \mathrm{w}, \theta)}{\nabla w} + \eta \tag{4}$$

With:

S = vector of cost shares;
η = under or over utilisation of resources;

Common practice is to select an appropriate mathematical function and appropriate statistical distributions for u (one-sided) and v (two-sided). An appropriate and very popular choice for the cost function is the translog function, which relates the logarithm of cost to the logarithms of services produced, resource prices and of all quadratic and cross-terms of services and resource prices (see e.g. Christensen et al. 1973). Many other functional forms are also available. In general, measurement error is specified by a normal distribution and the cost inefficiency component by a one-sided distribution such as half-normal, truncated normal or exponential. The parameters of the cost function and the statistical distribution are estimated by maximum likelihood methods (for the various estimation procedures see Kumbhakar and Lovell 2000). A general criticism is that in spite of the alleged flexibility of the functional specifications, it still is not flexible enough to model the complex cost structure. Particularly, in the case of a wide range and scope of the services delivered among firms, the cost structure of small and large firms or firms with a complete different service mix may differ to such an extent that it may be impossible to capture by a smooth function. It would make more sense to establish the cost structure locally. The parameters of the cost function that are estimated then depend on the data of firms that are assumed to have a similar cost structure. There are several ways to proceed here. In the case of very large number of observations Kernel estimators can be used. Put simply, kernel estimators calculate the average in a very close neighbourhood of the observation under investigation (reference observation), which can be regarded as a point estimate of the observation. Instead of averaging, one may also think of applying some local regression and use the

prediction of the regression as a point estimate. This would make more sense in a case when the neighbourhood can be tight as one may wish due to a lack of close neighbourhood observations. Another way to deal with this lack of close neighbourhood is by applying methods that put less weight on observations that are farther away from the observation under investigation such as weighted least squares. Then, the following set of equations of a cost function model has to be estimated:

$$\epsilon = weight^* \left[\ln(C) - c\left(\ln(y), \ln(w), \theta\right)\right] \tag{5}$$

$$\eta = weight^* \left[\mathrm{S} - \frac{\nabla c\left(\ln(y), \ln(w), \theta\right)}{\nabla \ln(w)} \right] \tag{6}$$

Note that ϵ reflects the composite error $u + v$. In the empirical application u will be estimated in an iterative procedure, which will be explained later on. Because of the singularity of the system one of the cost share equations must be dropped from the set of Eq. (6).

Since we are only interested in a local estimator of the production technology at a given observation i (= *1, .., I*), it suffices to use a first-order Taylor approximation at the given point. However, there is no objection whatsoever to using higher order expansions, except for the number of parameters to be estimated. Note that we only use the Taylor approximation for an estimate of the cost and the gradient of the cost at that particular point. The cost function can be written as:

$$\ln(C) = a_0 + \sum\nolimits_m^M b_m \ln\left(y_m\right) + \sum\nolimits_n^N c_n \ \ln\left(w_n\right) + \sum\nolimits_k^K d_k \ \ln\left(z_k\right) + h_1 time \tag{7}$$

With:

y_m = output m;
w_n = input price n;
z_k = environmental characteristic k;
$time$ = trend;
b_m, c_n, d_k, h_1 parameters to be estimated;

In addition, we also estimate the cost share equations simultaneously as:

$$sh_n = c_n \quad (n = 1, .., N) \tag{8}$$

With:

sh_n = cost share of input n;

The system of equations will be estimated with weighted nonlinear least squares. The weights are based on the distance of the reference observation to the other observations and to the frontier. The idea behind this approach is, to put it simply, that the estimates should be based on efficient neighbours whenever possible. The

extent to which this is possible is an empirical matter. The weight function, for instance, can be described as:

$$weight = eff^{*} \text{ norm}\left[\frac{dist}{k \bullet \sigma_{tot}}\right] \tag{9}$$

With:

eff = efficiency;
dist = distance to the observation under investigation;
σ_{tot} = standard deviation of distance measure;
k = scaling parameter;
norm(.) is the normal density function.

The cost efficiency score is derived from the cost efficiency component u in the cost function. We therefore need a point estimate of u. Here we use the conditional mode of the half-normal distribution, as suggested by Materov (1981). This estimator has the useful property that the mode is proportionally related to the composite error ϵ in the case of $\epsilon > 0$ (with the proportion being the ratio between the variance of the efficiency component and total variance) and equals zero in the case of $\epsilon \leq 0$. After each least squares estimation for observation i, we set $weight_i = \exp(-M_i)$ with M_i being the aforementioned conditional mode. Note that the conditional mode is being used as a measure to disentangle the composite error ϵ, as well as to set the weight variable.

In order to obtain a distance measure that does not depend on the unit of measurement, all variables are standardised on their means. Then the distance is measured by the Euclidean distance:

$$dist = \sqrt{\sum_{m}^{M} \left(y_m - y_m^{*}\right)^2} \tag{10}$$

With:

dist = average distance to the reference observation;
y_m^{*} = value of output m of the reference observation.
weight = weight attached to an observation;
σ_{tot}^2 = sum of variances of y_m;
k is a (fixed) parameter.

For each observation $i = 1, .., I$ we apply weighted least squares (WLS) and we preserve the error ϵ_i. After the I-th analysis a point estimate of u and *weight* can be established and the next iteration of I WLS-analyses can be conducted. This procedure is repeated until the differences in cost efficiency between successive iterations become very small.

***To summarise*:**
The procedure is conducted in several stages $s = 1, \ldots, S$ and stops at iteration S when the efficiency scores change less than a threshold value ($= 0.01$). At $s = 1$, the vector of weight parameters is set to 1.

At each stage, weighted least squares is applied to a cost function model for each DMU separately, consisting of a cost function and cost share equations. The weights are based on the distance (*dist*) between a DMU and the DMU under investigation (reference observation) and the cost efficiency (*eff*) of the DMU: The larger the Euclidean distance, the smaller the weight, and the larger the cost efficiency, the larger the weight.

Each separate WLS for a DMU provides an estimate of the cost efficiency parameter (u), which can be used in the next stage of the procedure to set the weight parameter. Note that the efficiency parameter varies only per stage and the *dist* parameter per WLS analysis.

At $s = S$, economic outcomes can be presented, such as scale elasticity, marginal cost, technical change, and cost efficiency scores.

3 Data

3.1 Production

The different types of schools in secondary education require different educational processes and consequently lead to different costs. For example, a teacher who teaches students in the final year of pre-academic education is generally more expensive than a teacher for students in the first year of vocational training. Therefore, production cannot be captured in a single number. Production indicators are based on the different types of education and grades. We therefore distinguish:

- Grade 1 and 2 of all types of education;
- Grade 3–4 vocational training;
- Grade 3–6 general higher and pre-academic.

Quality in education is generally difficult to measure. In order to take the quality of education into account, passes to next grades and examination results are included. The influence of the initial skills of pupils on quality measures are taken into account by including the so-called school recommendation at the start of a pupil's school career.

3.2 The Resources

The resources used can be divided into five categories or types of costs:

- Teaching personnel;
- Administrative personnel;

- Executive board and management;
- Housing (excluding rent);
- Material supplies.

We exclude capital cost because for most institutions, local government is responsible for providing the school buildings. Therefore, rent and amortisation of buildings are excluded in order to provide a meaningful comparison with institutions that own their school buildings.

3.3 Resource Prices

The relative prices of the staff categories differ by region and year. Averaging personnel costs per full-time equivalent over regions and years by a regression analysis provides a labour price for each staff category for each region in a particular year.

The prices for housing and material are assumed to be equal for all educational institutions and thus only vary over the years. Since housing costs are restricted to building-related costs such as energy and cleaning, the energy price indices of Statistics Netherlands are used for housing costs. For material costs, the consumer price index of Statistics Netherlands is used.

3.4 Data Resources, Data Checks and Manipulations

For the analyses, we used different databases. The number of pupils was taken from the public files of the Office of Education (DUO) and the Ministry of Education, Culture and Science. The numbers on education returns were supplied by the Education Inspectorate. The staff numbers and salary data were also provided by DUO. Finally, the price development of energy and consumer goods and services was collected by Statistics Netherlands. The period for which all the necessary data are available is 2007–2010.

3.5 Data Checks and Manipulations

We applied a number of checks and manipulations to these data (for details, see Urlings and Blank 2012). A statistical description of the data for the year 2010 is given in Table 1.

Table 1 Statistical description variables in analysis, 2010

Variable	Mean	Std. Error	Minimum	Maximum
Grades 1-2[a]	1148.5	821.5	185.1	5783.3
Vocational training grades 3-4[a]	515.8	375.7	76.5	2501.3
General education grades 4-6[a]	650.7	491.4	93.1	3095.3
Total cost (x € 1000)	19,179	14,358	5100	105,563
Cost share board/management	0.05	0.02	0.00	0.16
Cost share administrative personnel	0.09	0.04	0.00	0.24
Cost share teaching personnel	0.65	0.05	0.44	0.81
Cost share housing	0.07	0.03	0.02	0.28
Cost share material supplies	0.14	0.03	0.06	0.29
Price management (€)	100014.3	4960.3	88393.0	110339.0
Price administrative personnel (€)	46388.9	3872.6	37756.0	52967.0
Price board/management (€)	65278.5	3937.8	57210.0	74661.0
Price housing (€)	358.5	14.4	342.0	380.5
Price material supplies (2007 = 100)	104.6	1.8	101.6	106.7

[a]Corrected with pass rate

3.6 Secondary Education Statistics

In 2010, the average secondary school in the Netherlands had 3300 pupils. Of these, 38% were in the first two grades, 19% in junior vocational education, 35% in senior general secondary education or pre-university education, and 8% in other education (practical education, primary education or senior vocational education). The costs can be divided across five categories:

- teaching staff (65%);
- administrative staff (9%);
- management (5%);
- accommodation (6%);
- material supplies (15%).

There is a strong variation in the scale of the educational institutions. Half of the educational institutions have fewer than 2100 pupils and costs of under 17.5 million euros. The largest educational institution has over 62,000 pupils and costs totalling 482 million euros.

4 Results

The outcomes are presented as graphs. Figures 1, 2 and 3 present the marginal costs of the different types of pupils who passed. The marginal cost gives a first indication of the plausibility of the estimates. The marginal costs of an undergraduate pupil follow more or less a normal distribution of €8000 with a limited variance. Marginal costs of pupils in vocational training are higher, distributed around € 10,000. This

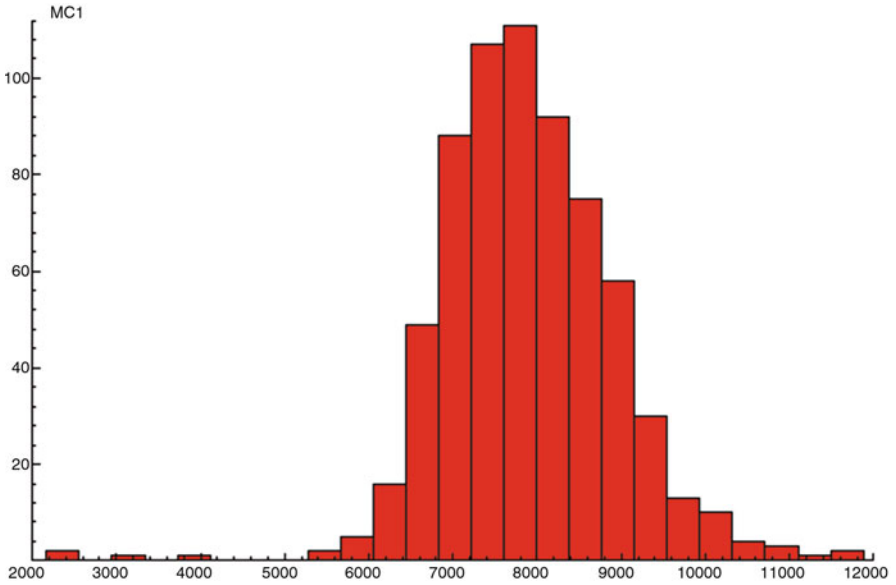


Fig. 1 Estimated marginal costs of undergraduate pupils (corrected for passes)

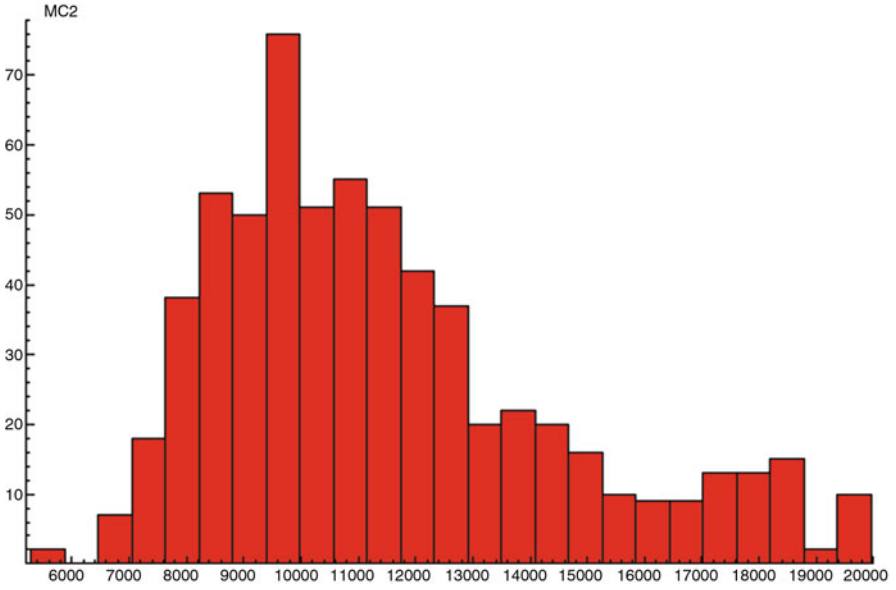


Fig. 2 Estimated marginal costs of pupils in vocational training (corrected for passes)

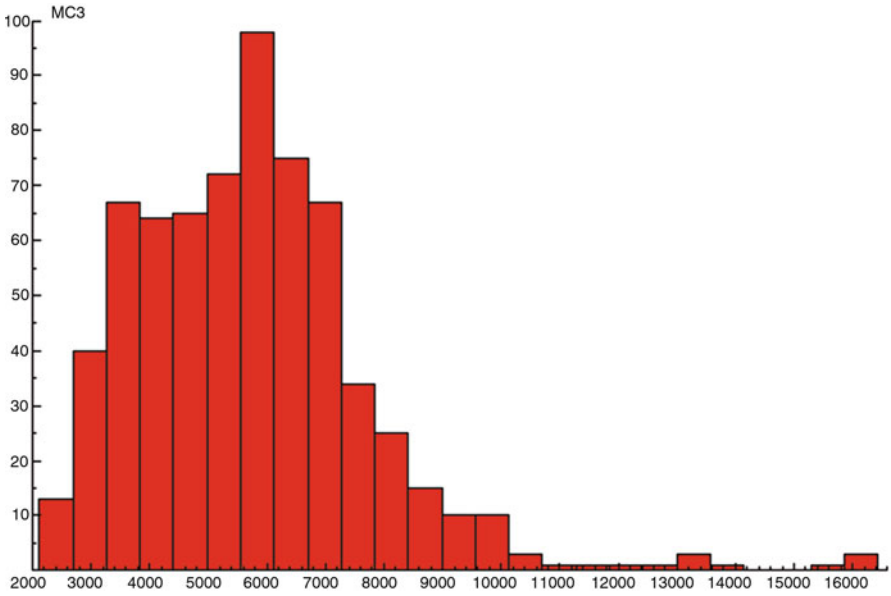


Fig. 3 Estimated marginal costs of pupils in general education (corrected for passes)

distribution is skewed to the right. A number of schools tend to have higher marginal costs than the modus. The marginal costs of pupils in general education (Fig. 3) has a distribution around € 6000. This distribution is skewed to the left, indicating that there are relatively more schools with lower marginal costs than schools with higher marginal costs. The outcomes make sense, since it is known that vocational training is more expensive than general education due to higher material costs (machinery, etc.). General education is less expensive than undergraduate education due to the substantially lower number of teaching hours in the graduate phase of education.

Figure 4 represents the estimated cost flexibility of each school board (the red + signs). Each school board is reflected by its size, expressed in terms of a number of times the average size. So two, for instance, reflects a school board that is twice the size of the average school board. The green and purple lines represent the lower and upper bounds of the 95% confidence interval. Outcomes less than one indicate economies of scale, and outcomes greater than one indicate diseconomies of scale. From Fig. 4 we can conclude that school boards less than two times the size of the average school board face economies of scale, whereas school boards greater than three times the average size face diseconomies of scale. The optimum size (with neither economies nor diseconomies of scale) lies between two and three times the average size.

Figure 5 represents the distribution of the cost efficiency scores. It shows that the majority of the school boards are cost efficient or close to cost efficient. This is due to the fact that school boards are only regarded as suitable references when they have a broadly similar production profile. Observations with deviated production profiles receive a low weight in the estimation procedure. Sensitivity analysis based on local

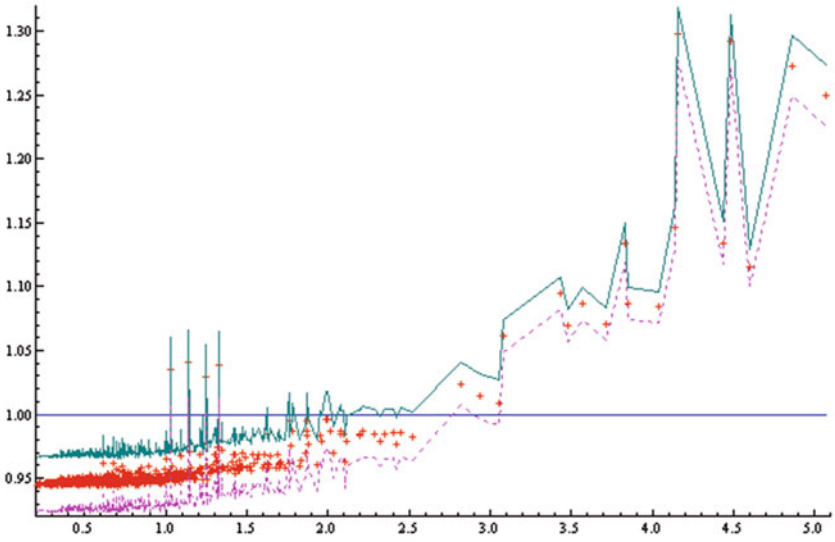


Fig. 4 Economies of scale (cost flexibility with 95% CI)

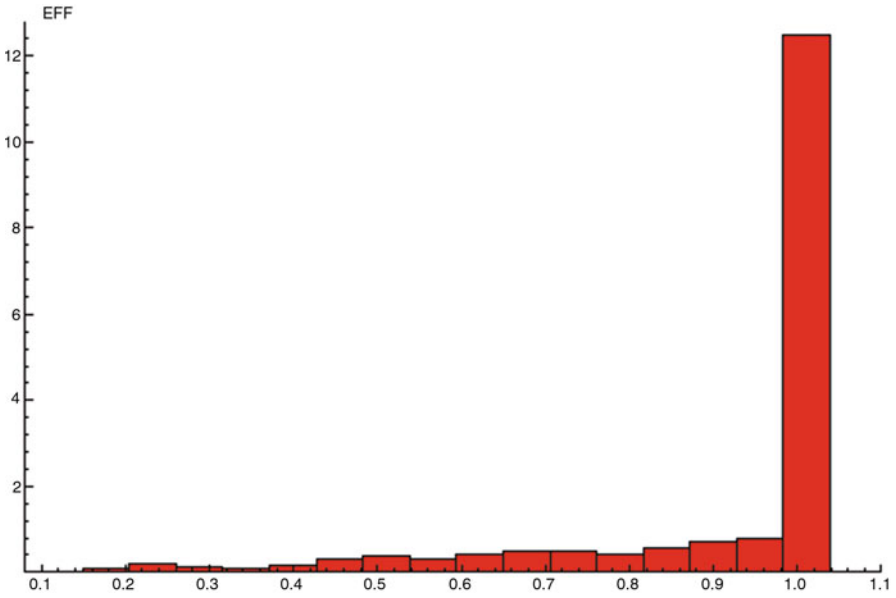


Fig. 5 Cost efficiency scores

estimation with a different weighting scheme may result in a different cost efficiency pattern. However, it shows that estimation with different weighting schemes leads to almost identical distributions of efficiency scores.

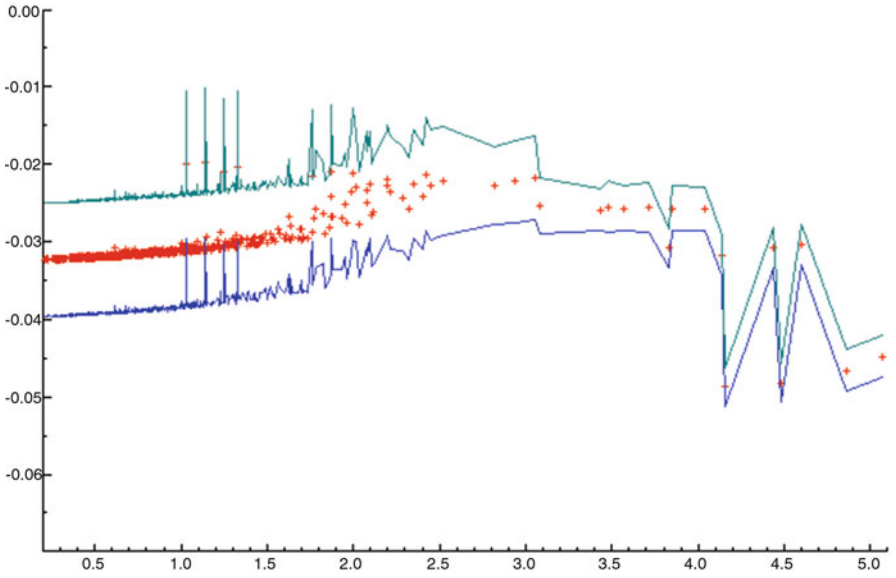


Fig. 6 Technical change (with 95% CI)

Figure 6 displays technical change. The mean is represented by the red dots, while the 95% confidence intervals are represented by the upper (green) and lower bounds (blue). It shows that technical change is significantly negative for all schools, implying a loss of productivity of about 3% annually. It also shows that for a few larger school boards, technical change is even more negative. This may seem to be an incredible outcome. However, education is certainly not an exception. In recent years, an extensive research programme on public sector productivity has been carried out in the Netherlands. A substantial number of studies in different sectors show negative technical change (see e.g. Blank and Heezik van 2017; Blank and van Heezik 2015; Blank et al. 2016). An intensifying administrative burden, obligatory protocols and systems, and other bureaucratic tendencies may explain this negative development.

Additionally, we checked the outcomes on the conditional mean of the cost shares. They show a very consistent pattern with respect to size.

We also conducted some sensitivity analyses by varying the scaling parameter (k parameter in Eq. 9). The scaling parameter refers to the weighting scheme for conducting WLS. The weighting scheme follows a normal distribution with a variance parameter depending on the (Euclidean) distance of the service variables between the reference observation and other observations. Eq. (9) implies that for large k the weight distribution will be flatter than for small k. We calculated the outcomes for different k (= 0.25, 0.50 and 1.00) and tested the effect on the outcome. Table 2 summarises the outcomes by presenting a test of the mean difference in the cost efficiency scores, the estimates of technical change and the estimated cost flexibilities.

Table 2 Test results of mean differences for varying weighting schemes ($k = 0.25$, 0.5 and 1)

Test of mean difference	Mean	T-test
Cost efficiency $k = 1$vs $k = 0.5$	−0.001	1.58
Cost efficiency $k = 0.5$ vs $k = 0.25$	−0.002	−3.34
Technical change $k = 1$ vs $k = 0.5$	−0.000	−1.91
Technical change $k = 0.5$ vs $k = 0.25$	−0.001	−16.30
Cost flexibility $k = 1$vs $k = 0.5$	0.002	0.67
Cost flexibility $k = 0.5$ vs $k = 0.25$	−0.008	−8.05

From Table 2 we conclude that the differences between the different (point) estimators are relatively low. In the case of the cost efficiency scores, a flatter weighting scheme (corresponding to a larger k) corresponds to lower cost efficiency scores. This makes sense, since observations that are far away from the reference point still have a substantial influence (in contrast with the case where these observations get a small weight with small k). The differences, however, are very small and in one case not significantly different from zero. Note here that cost efficiency scores are about 95% on average. For technical change, the differences are also very small, in spite of the fact that the mean differences between technical change for $k = 0.5$ and $k = 0.25$ are significant. The same holds for cost flexibility. The differences are less than 1%. Flatter weighting schemes ($k = 0.5$ versus $k = 0.25$) correspond to a slight decrease in cost flexibility, implying that economies of scale are stronger. Furthermore, it is striking that, although the differences are very small, all differences are significant at the 5%-level with respect to the test $k = 0.5$ versus $k = 0.25$.

5 Conclusion

In this paper, we have presented a method for (cost) efficiency measurement that is based on the idea of local estimation in several stages. The new element in the method is that it also includes the information from the cost share equations and includes cost efficiency in the distance measure. The method is based on weighted least squares, where weights depend on the distance of an observation to all other observations (more or less lookalikes) and on the distance to the cost frontier. The latter is derived from a former stage and implemented in a next stage analysis. An application to a data set of Dutch school boards in secondary education shows that it works well. The approach produces a number of reliable estimates. It also shows a variation in outcomes that would be hard to cover with, for instance, traditional procedures such as SFA on a translog cost function. It therefore seems that the proposed approach could be an interesting alternative to standard frontier techniques. This approach adds more flexibility to the modelling of production technology.

Nevertheless, a number of issues still need to be addressed. The set of weights is based on a distance measure and the cost efficiency score. For the distance measure, a traditional Euclidean measure is used, whereas cost efficiency scores are assumed to be directly related to the estimated errors in the first stage of the procedure. Alternative distance measures and cost efficiency estimates should be used to indicate the sensitivity to these assumptions. Some sensitivity analyses have been carried out here. The application that has been demonstrated here indicates that varying the weighting schemes has only limited effects on the outcomes. However, more research on the effect of alternative distance and cost efficiency measures is required.

References

Aigner, D., Lovell, C. A. K., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, *6*(1), 21–37. https://doi.org/10.1016/0304-4076(77)90052-5.

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale efficiencies in data envelopment analysis. *Management Science, 30*(9), 1078–1092.

Blank, J. L. T. (2000). *Public provision and performance: Contributions from efficiency and productivity measurement*. Amsterdam: Elsevier.

Blank, J. L. T., & Eggink, E. (2004). The decomposition of cost efficiency: An empirical application of the shadow cost function model to Dutch general hospitals. *Health Care Management Science, 7*, 79–88.

Blank, J. L. T., & Heezik van, A. A. S. (2017). Productiviteit van overheidsbeleid, deel III: de Nederlandse veiligheid en justitie, 1980–2014. forthcoming.

Blank, J. L. T., & Meesters, A. (2012). Iteratively weighted least squares on stochastic frontier estimation. Applied to the Dutch hospital industry. In R. Banker, A. Emrouznejad, A. L. Miranda Lopez, & M. R. de Almeida (Eds.), *Data envelopment analysis: Theory and applications. Proceedings of the 10th international conference on DEA* (pp. 167–185). Natal.

Blank, J. L. T., & van Heezik, A. A. S. (2015). *Productiviteit van overheidsbeleid, deel I: het Nederlandse onderwijs, 1980–2012*. Den Haag/Delft: Eburon.

Blank, J. L. T., van Heezik, A. A. S., & Niaounakis, T. K. (2016). *Productiviteit van overheidsbeleid, deel II: de Nederlandse zorg, 1980–2013*. Den Haag/Delft: Eburon.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*(6), 429–444. https://doi.org/10.1016/0377-2217(78)90138-8.

Christensen, L. R., Jorgenson, D. W., & Lau, L. J. (1973). Transcendental logarithmic production Frontiers. *The Review of Economics and Statistics, 55*(1), 28–45. Retrieved from http://www.jstor.org/stable/1927992.

Coelli, T. J., Rao, D. S. P., O'Donnell, C. J., & Battese, G. E. (2005). *An introduction to efficiency and productivity analysis* (2nd ed.). New York: Springer.

Debreu, G. (1951). The coefficient of resource utilization. *Econometrica, 19*(3), 273–292. Retrieved from http://www.jstor.org/stable/1906814.

Fan, Y., Li, Q., & Weersink, A. (1996). Semiparametric estimation of stochastic production frontier models. *Journal of Business & Economic Statistics, 14*(4), 460–468

Färe, R., Grosskopf, S., & Lovell, C. A. K. (1986). Scale economies and duality. *Journal of Economics-Zeitschrift Fur Nationalokonomie, 46*(2), 175–182. https://doi.org/10.1007/bf01229228.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General), 120*(3), 253–290. Retrieved from http://www.jstor.org/stable/2343100.

Fried, H. O., Lovell, C. A. K., & Schmidt, S. S. (2008). *The measurement of productive efficiency and productivity growth*. New York: Oxford University Press.

Johnson, A., & Kuosmanen, T. (2015). An introduction to CNLS and StoNED methods for efficiency analysis: Economic insights and computational aspects. *Benchmarking for Performance Evaluation.*, 1–80, Retrieved from http://link.springer.com/chapter/10.1007/978-81-322-2253-8_3.

Kumbhakar, S. C. (1997). Modelling allocative inefficiency in a Translog cost function and cost share equations. *Journal of Econometrics, 76*, 351–356.

Kumbhakar, S. C., & Lovell, C. (2000). *Stochastic frontier analysis*. New York: Cambridge University Press.

Kumbhakar, S. C., & Tsionas, E. G. (2005). The joint measurement of technical and allocative inefficiencies: An application of Bayesian inference in nonlinear random-effects models. *Journal of the American Statistical Association, 100*(471), 736–747.

Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. *The Econometrics Journal, 8*(2), 308–325. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1368-423X.2008.00239.x/full.

Maietta, O. W. (2002). The effect of the normalisation of the shadow price vector on the cost function estimation. *Economics Letters, 77*, 381–385.

Materov, I. S. (1981). On full identification of the stochastic production frontier model. *Ekonomika I Matematicheskie Metody, 17*, 784–788.

Meeusen, W., & Van den Broeck, J. (1977). Efficiency estimation from cobb-Douglas production functions with composed error. *International Economic Review, 8*, 435–444.

Parmeter, C., & Kumbhakar, S. (2014). *Efficiency analysis: A primer on recent advances*. New York: Miami.

Ray, S. C., Kumbhakar, S. C., & Dua, P. (2015). *Benchmarking for performance evaluation: A production frontier approach. Benchmarking for performance evaluation: A production frontier approach*. New Delhi: Springer India. https://doi.org/10.1007/978-81-322-2253-8.

Shephard, R. W. (1953). *Theory of cost and production functions*. (P. U. Press, Ed.). Princeton: Princeton University Press.

Urlings, T., & Blank, J. L. T. (2012). *Benchmark bedrijfsvoering voortgezet onderwijs. IPSE Studies Research Reeks*. Delft: TU Delft, IPSE Studies.

Aggregate Productivity and Productivity of the Aggregate: Connecting the Bottom-Up and Top-Down Approaches

Bert M. Balk

Abstract Productivity analysis is carried out at various levels of aggregation. In microdata studies the emphasis is on individual firms (or plants), whereas in sectoral studies it is on (groupings of) industries. An industry is an ensemble of individual firms (decision making units) that may or may not interact with each other. In National Accounts terms this is symbolized by the fact that industry (aggregate) nominal value added is the simple sum of firm-specific nominal value added. From this viewpoint it is natural to expect there to be a relation between industry productivity and the firm-specific productivities. Yet, microdata researchers do not appear to pay much attention to the interpretation of the weighted means of firm-specific productivities they employ in their analyses. In this paper the consequences of this are explored, based on a review of the literature.

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

The point of departure of this paper is that aggregate productivity should be interpreted as productivity of the aggregate. It is shown that this implies restrictive

This paper draws from an extended version available at SSRN: http://ssrn.com/abstract=2585452. Presentations took place at the North American Productivity Workshop IX, 15–18 June 2016, Québec City, and the 34th General Conference of the International Association for Research in Income and Wealth, 21–27 August 2016, Dresden. Susanto Basu informed me that he had once written a paper with a virtually identical title. On inspection this turned out to be an embryonic version of Basu and Fernald (2002). Though related, my paper has a different focus.

B.M. Balk (✉)
Rotterdam School of Management, Erasmus University, Rotterdam, Netherlands
e-mail: bbalk@rsm.nl

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_5

relations between the productivity measure, the set of weights, and the type of mean employed. For instance, value-added based total factor productivities and output based weights require a harmonic mean, if additivity is assumed.

Keywords Producer · Productivity · Aggregation · Bottom-up approach · Top-down approach · Index number theory

JEL code: C43, D24, O47

1 Introduction

In the first article of this series, Balk (2010), I considered productivity measurement for a single, consolidated production unit. In terms of levels, productivity is defined as real output divided by real input. Real output or input means nominal output or input deflated by some output- or input-specific price index, respectively. For the production unit considered, productivity change (through time) can then be measured as a difference or a ratio of productivities. In the latter case it appears that productivity change can also be defined directly as output quantity index divided by input quantity index.

The choice of the output and input concepts appears to be critical. Three main models can be distinguished: KLEMS-Y, KL-VA, and K-CF. Taking the composition of capital input cost into account, as set out in Balk (2011), two more models can be added, namely KL-NVA and K-NCF. Assuming profit (defined as revenue minus total cost) to be equal to zero, or, what amounts to the same, replacing an exogenous interest rate by an endogenous rate, multiplies the number of models by two. And the introduction of a capital utilization rate further complicates the picture. Thus, there is a lot of choice here, with not unimportant empirical consequences, as illustrated by Vancauteren et al. (2012).

Production units exist at various levels of aggregation. We see plants, enterprises, industries, countries, to name just some types of production units materializing in analyses of productivity change. Usually such units appear, more or less naturally, arranged into higher level aggregates: a number of plants belonging to the same enterprise; a certain type of enterprises defining an industry; a number of industries defining the 'measurable' part of a national economy; national economies making up the world economy. It is not difficult to perceive several sorts of hierarchy here.

As in any of these situations the structure is the same – there is an ensemble of production units, and the ensemble itself may or may not be considered as a higher level production unit –, it is interesting to study the relation between aggregate productivity (change) and productivity (change) of the aggregate.

There are basically two approaches here. Balk (2016) reviews and discusses the so-called *bottom-up* approach, the approach that takes an ensemble of individual production units as the fundamental frame of reference. The *top-down* approach is the subject of three other papers, namely Balk (2014) plus Dumagan and Balk (2016) on labour productivity, and Balk (2015) on total factor productivity.

The present paper investigates the connection between the two approaches, bottom-up and top-down. Characteristic of the approach taken in this paper is that aggregate productivity (change) should be interpreted as productivity (change) of the aggregate. It will be shown that this implies restrictive relations between the productivity measures involved, including the weights of the individual production units, and the type of mean employed. For instance, it appears that, assuming additivity, value-added based total factor productivities and output based weights require a harmonic mean.

The order of this paper is as follows. Section 2 reviews basic accounting relations. Section 3 defines the problem. Sections 4 and 5 consider value-added based total factor productivity and labour productivity, respectively. Section 6 considers gross-output based productivity. Section 7 concludes.

2 Accounting Framework

We consider[1] an ensemble (or set) $\mathcal{K}^t$ of consolidated production units,[2] 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

$$C^{kt}_{KL} + C^{kt}_{EMS} + \Pi^{kt} = R^{kt} \; (k \in \mathcal{K}^t), \tag{1}$$

where C^{kt}_{KL} denotes the primary input cost, C^{kt}_{EMS} 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^{kt}_{EMS} 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 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^{kt}_{K} 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

[1] This section has been copied from Balk (2016). Though the time dimension does not play an explicit role in the present paper, the notation is retained for consistency.

[2] "Consolidated" means that intra-unit deliveries are netted out. At the industry level, 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. In terms of variables to be defined below, consolidation means that $C^{kkt}_{EMS} = R^{kkt} = 0$.

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

$$C_{KL}^{kt} + \Pi^{kt} = R^{kt} - C_{EMS}^{kt} \equiv VA^{kt} \; (k \in \mathcal{K}^t), \tag{2}$$

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 (1) 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,

$$C_{EMS}^{kt} = \sum_{k' \in \mathcal{K}^t, k' \neq k} C_{EMS}^{k'kt} + C_{EMS}^{ekt}, \tag{3}$$

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 $\mathcal{K}^t$. Similarly,

$$R^{kt} = \sum_{k' \in \mathcal{K}^t, k' \neq k} R^{kk't} + R^{ket}, \tag{4}$$

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 (1) then delivers

$$\sum_{k \in \mathcal{K}^t} C_{KL}^{kt} + \sum_{k \in \mathcal{K}^t} \sum_{k' \in \mathcal{K}^t, k' \neq k} C_{EMS}^{k'kt} + \sum_{k \in \mathcal{K}^t} C_{EMS}^{ekt} + \sum_{k \in \mathcal{K}^t} \Pi^{kt} =$$

$$\sum_{k \in \mathcal{K}^t} \sum_{k' \in \mathcal{K}^t, k' \neq k} R^{kk't} + \sum_{k \in \mathcal{K}^t} R^{ket}. \tag{5}$$

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[3]

$$\sum_{k \in \mathcal{K}^t} C_{KL}^{kt} + \sum_{k \in \mathcal{K}^t} C_{EMS}^{ekt} + \sum_{k \in \mathcal{K}^t} \Pi^{kt} = \sum_{k \in \mathcal{K}^t} R^{ket}. \tag{6}$$

[3] See Balk (2015, footnote 2) for the treatment of net taxes on intermediates.

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 (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

$$\sum_{k\in\mathcal{K}^t} C_{KL}^{kt} + \sum_{k\in\mathcal{K}^t} \Pi^{kt} = \sum_{k\in\mathcal{K}^t} R^{ket} - \sum_{k\in\mathcal{K}^t} C_{EMS}^{ekt}, \tag{7}$$

which can be written as

$$C_{KL}^{\mathcal{K}^t t} + \Pi^{\mathcal{K}^t t} = R^{\mathcal{K}^t t} - C_{EMS}^{\mathcal{K}^t t} \equiv VA^{\mathcal{K}^t t}. \tag{8}$$

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

$$VA^{\mathcal{K}^t t} = \sum_{k\in\mathcal{K}^t} VA^{kt}. \tag{9}$$

The structural similarity between expressions (2) and (8), together with the additive relations between all their 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 soon discover, however, that the bottom-up approach basically neglects this framework.

3 Bottom-Up and Top-Down Approaches Connected

Let the productivity level[4] of unit k at period t be denoted by $PROD^{kt}$. The generic definition here employed is: real output divided by real input. Output can be measured as revenue (also called 'gross output') (R^{kt}) or as value added (VA^{kt}). Input can be measured as total cost ($C_{KLEMS}^{kt} \equiv C_{KL}^{kt} + C_{EMS}^{kt}$), as primary input cost (C_{KL}^{kt}), as labour input cost (C_{L}^{kt}), or as total labour quantity (L^{kt}, where a common unit is used for the various types of labour). In all these cases, 'real' means nominal deflated by some price index, which may or may not be specific for each production unit. It is supposed that the reference period b, that is, the period for which the price index equals 1 by definition, is the same for all the units.

Each production unit comes with some measure of relative size (or, importance) in the form of a weight θ^{kt}. For each period these weights usually but not necessarily add up to 1.

[4] On the relation between levels and indices, see Balk (2016, 21–28).

The question as to 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 the weight 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.[5] 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 definitions introduced in Balk (2016), *aggregate productivity* is a weighted 'mean' of the individual productivities

$$PROD^t \equiv M(\theta^{kt}, PROD^{kt}; k \in \mathcal{K}^t), \tag{10}$$

where the 'mean' $M(.)$ 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 total factor productivity, labour productivity or simple labour productivity, as defined in Balk (2016), or gross-output based total factor productivity or simple labour productivity, to be defined in this paper.

Microdata studies, where the production units considered are plants or enterprises, then concentrate on the distributional characteristics of the (large) set of individual productivities $PROD^{kt}$, the development over time of aggregate productivity $PROD^t$, and the decomposition of this development with respect to several types of firms.

Sectoral studies, where the production units considered are industries (according to some national or international classification), are usually interested in industry-specific productivity change and its components, such as capital deepening and labour-composition change. The number of industries distinguished is generally so small that separate attention can be devoted to each specific case.

In both situations the ensemble $\mathcal{K}^t$ itself can be considered as a (consolidated) higher level production unit. Using the same definitions, its productivity $PROD^{\mathcal{K}^t t}$ can be calculated. In general it will then turn out, explicitly or implicitly, that the *productivity of the aggregate*, $PROD^{\mathcal{K}^t t}$, is unequal to aggregate productivity, $PROD^t$, as defined above.[6]

[5] de Loecker and Konings (2006) noted that there is no clear consensus on the appropriate weights (shares) that should be used. In their own work they used employment based shares $L^{kt}/\sum_k L^{kt}$ to weigh value-added based total factor productivity indices $Q^k_{VA}(t,b)/Q^k_{KL}(t,b)$. We will return to this example.

[6] $PROD^t$ can be considered as a 2-stage aggregation procedure: first $PROD^{kt}$ aggregates over basic inputs and outputs per production unit k, and then $PROD^t$ aggregates over all the units $k \in \mathcal{K}^t$. $PROD^{\mathcal{K}^t t}$ can be considered as a 1-stage aggregate of the same basic inputs and outputs. See Diewert (1980, 495–498) for a similar discussion in terms of variable profit (or, value added) functions and technological change (assuming continuous time and differentiability), and the PPI

Microdata analysis is usually not interested in the productivity of the aggregate. As a consequence the problem of the choice of weights and type of mean arises. Sectoral analysis usually does show productivity change of the aggregate (e.g., the economy) alongside productivity change of the component industries, however without an explicit discussion of their relationship. If there is some comparison of aggregate productivity change and productivity change of the aggregate at all, then their difference is classified as an "unexplained residual".

In this paper we will ask whether it is possible to find a set of weights and a type of 'mean' such that

$$PROD^{t} = PROD^{\mathcal{K}^t t}; \tag{11}$$

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

As we know, there are a number of options here. We start with the case where $PROD^{kt}$ and $PROD^{\mathcal{K}^t t}$ is value-added based total factor productivity. Next we consider value-added based labour productivity. Finally we turn to gross-output based labour and total factor productivity respectively.

4 Value-Added Based Total Factor Productivity

The top-down approach starts with the adding-up relation (9). This relation tells us that nominal value added of the ensemble $\mathcal{K}^t$ is the sum of nominal value added of the individual production units k making up this ensemble. Next it is important to recall that the KL-VA accounting identities of the individual units, given by expression (2), are structurally identical to the KL-VA accounting identity of the ensemble (8). This means that we can treat the ensemble as a higher level production unit, and that all the definitions of indices and levels can be applied to the individual units and the ensemble in the same way.

Real value added, $RVA^{k}(t, b)$, is nominal value added, VA^{kt}, divided (or, deflated) by some price index with reference period b, $P^{k}_{VA}(t, b)$. Rewriting this definition gives

$$VA^{kt} = P^{k}_{VA}(t, b) RVA^{k}(t, b) \ (k \in \mathcal{K}^t). \tag{12}$$

Nominal value added is here decomposed into a price component and a quantity component. For the ensemble we have similarly

$$VA^{\mathcal{K}^t t} = P^{\mathcal{K}^t}_{VA}(t, b) RVA^{\mathcal{K}^t}(t, b), \tag{13}$$

Manual (2004, Chapter 18) for the cases of revenue, intermediate-input-cost, and value-added based price indices. Notice the double role of the variable t in $PROD^{\mathcal{K}^t t}$.

where $P_{VA}^{\mathcal{K}^t}(t,b)$ is a value-added based price index for the ensemble $\mathcal{K}^t$ for period t relative to the reference period b. This index is supposed to be estimated from a sample of enterprises and products.

Substituting expressions (12) and (13) into expression (9) and dividing both sides by the price index $P_{VA}^{\mathcal{K}^t}(t,b)$ delivers a relation between real value added of the ensemble and real value added of the individual units,

$$RVA^{\mathcal{K}^t}(t,b) = \sum_{k\in\mathcal{K}^t} \frac{P_{VA}^k(t,b)}{P_{VA}^{\mathcal{K}^t}(t,b)} RVA^k(t,b). \tag{14}$$

It is important to observe that, unlike nominal value added – see expression (9) – real value added in general is not additive; that is, $RVA^{\mathcal{K}^t}(t,b) \neq \sum_{k\in\mathcal{K}^t} RVA^k(t,b)$.

For any individual production unit, real primary input is defined by

$$X_{KL}^k(t,b) \equiv C_{KL}^{kt}/P_{KL}^k(t,b) \ (k \in \mathcal{K}^t) \tag{15}$$

where $P_{KL}^k(t,b)$ is a suitable deflator for the primary input cost of production unit k. For the ensemble the corresponding definition reads

$$X_{KL}^{\mathcal{K}^t}(t,b) \equiv C_{KL}^{\mathcal{K}^t t}/P_{KL}^{\mathcal{K}^t}(t,b), \tag{16}$$

where $C_{KL}^{\mathcal{K}^t t} \equiv \sum_{k\in\mathcal{K}^t} C_{KL}^{kt}$ and $P_{KL}^{\mathcal{K}^t}(t,b)$ is a suitable deflator for the primary input cost of the ensemble $\mathcal{K}^t$. Now, dividing both sides of expression (14) by $X_{KL}^{\mathcal{K}^t}(t,b)$ and inserting at the right-hand side $X_{KL}^k(t,b)/X_{KL}^k(t,b) = 1$ $(k \in \mathcal{K}^t)$, one obtains

$$\frac{RVA^{\mathcal{K}^t}(t,b)}{X_{KL}^{\mathcal{K}^t}(t,b)} = \sum_{k\in\mathcal{K}^t} \frac{P_{VA}^k(t,b)}{P_{VA}^{\mathcal{K}^t}(t,b)} \frac{X_{KL}^k(t,b)}{X_{KL}^{\mathcal{K}^t}(t,b)} \frac{RVA^k(t,b)}{X_{KL}^k(t,b)}. \tag{17}$$

At both sides of this identity we see value-added based total factor productivity, as introduced by Balk (2016), for the aggregate and the individual production units, respectively. Thus expression (17) can be written as

$$TFPROD_{VA}^{\mathcal{K}^t}(t,b) = \sum_{k\in\mathcal{K}^t} \frac{P_{VA}^k(t,b)}{P_{VA}^{\mathcal{K}^t}(t,b)} \frac{X_{KL}^k(t,b)}{X_{KL}^{\mathcal{K}^t}(t,b)} TFPROD_{VA}^k(t,b). \tag{18}$$

This is our desired result. It means that if $PROD^{kt}$ is defined as value-added based total factor productivity $TFPROD_{VA}^k(t,b)$, then the appropriate weights are given by

$$\phi^{kt} \equiv \frac{P_{VA}^k(t,b)}{P_{VA}^{\mathcal{K}^t}(t,b)} \frac{X_{KL}^k(t,b)}{X_{KL}^{\mathcal{K}^t}(t,b)} \ (k \in \mathcal{K}^t). \tag{19}$$

When these weights are used, then aggregate productivity $\sum_{k\in\mathcal{K}^t} \phi^{kt} PROD^{kt}$ can be interpreted as the value-added based total factor productivity of the ensemble, considered as a higher-level production unit. Notice that the weights ϕ^{kt} $(k \in \mathcal{K}^t)$

do not necessarily add up to 1. Thus, though expression (18) is a weighted sum of individual productivities, it is not a genuine mean.[7]

There is, however, another way of looking at expression (18). To see this, notice that $(P^k_{VA}(t,b)/P^{\mathcal{K}^t}_{VA}(t,b))TFPROD^k_{VA}(t,b)$ is so-called revenue total factor productivity; that is, the result of deflating VA^{kt} not by its unit-k-specific deflator $P^k_{VA}(t,b)$ but by the ensemble-specific deflator $P^{\mathcal{K}^t}_{VA}(t,b)$. Weighing these revenue total factor productivities by real input shares $X^k_{KL}(t,b)/X^{\mathcal{K}^t}_{KL}(t,b)$ then delivers aggregate total factor productivity. Notice that these real input shares also do not necessarily add up to 1.

Another interesting viewpoint[8] emerges when the weights ϕ^{kt} are decomposed as

$$\phi^{kt} \equiv \frac{P^k_{VA}(t,b)X^k_{KL}(t,b)}{\sum_{k\in\mathcal{K}^t} P^k_{VA}(t,b)X^k_{KL}(t,b)} \frac{\sum_{k\in\mathcal{K}^t} P^k_{VA}(t,b)X^k_{KL}(t,b)}{P^{\mathcal{K}^t}_{VA}(t,b)X^{\mathcal{K}^t}_{KL}(t,b)} \ (k \in \mathcal{K}^t). \tag{20}$$

The first factor at the right-hand side is a share, adding up to 1 when summed over all $k \in \mathcal{K}^t$, whereas the second factor can be considered as an adjustment factor common to all the individual productivities $TFPROD^k_{VA}(t,b)$. Then expression (18) is a weighted (arithmetic) mean of adjusted individual productivities.

Expression (18) as a relation between aggregate and individual productivities is, however, not unique. To see this, instead of the adding-up relation for value added (9), we consider the adding-up relation for primary input cost,

$$C^{\mathcal{K}^t t}_{KL} = \sum_{k\in\mathcal{K}^t} C^{kt}_{KL}. \tag{21}$$

Employing definitions (15) and (16), expression (21) can be rewritten as

$$X^{\mathcal{K}^t}_{KL}(t,b) = \sum_{k\in\mathcal{K}^t} \frac{P^k_{KL}(t,b)}{P^{\mathcal{K}^t}_{KL}(t,b)} X^k_{KL}(t,b). \tag{22}$$

It is not unimportant to observe that, unlike nominal primary input cost, real primary input appears generally to be non-additive; that is, $X^{\mathcal{K}^t}_{KL}(t,b) \neq \sum_{k\in\mathcal{K}^t} X^k_{KL}(t,b)$.

[7] Expression (18) is the model underlying GEAD-TFP as implemented by Calver and Murray (2016). Stated in our notation, instead of the right-hand side of expression (18) Basu and Fernald (2002) consider

$$\sum_{k\in\mathcal{K}^t} \frac{VA^{kt}}{VA^{\mathcal{K}^t t}} TFPROD^k_{VA}(t,b);$$

that is, mean value-added based total factor productivity where the weights are nominal value-added shares. This, then, cannot be interpreted as value-added based total factor productivity of the ensemble, unless special conditions apply.

[8] This paragraph has been inserted at the suggestion of a referee.

Individual and aggregate real value added were defined by expressions (12) and (13) respectively. Now, dividing both sides of expression (22) by $RVA^{\mathcal{K}^t}(t,b)$ and inserting at the right-hand side $RVA^k(t,b)/RVA^k(t,b) = 1$ $(k \in \mathcal{K}^t)$, one obtains

$$\frac{X_{KL}^{\mathcal{K}^t}(t,b)}{RVA^{\mathcal{K}^t}(t,b)} = \sum_{k \in \mathcal{K}^t} \frac{P_{KL}^k(t,b)}{P_{KL}^{\mathcal{K}^t}(t,b)} \frac{RVA^k(t,b)}{RVA^{\mathcal{K}^t}(t,b)} \frac{X_{KL}^k(t,b)}{RVA^k(t,b)}. \tag{23}$$

Again employing the definition of value-added based total factor productivity, expression (23) can be written as

$$\left(TFPROD_{VA}^{\mathcal{K}^t}(t,b)\right)^{-1} = \sum_{k \in \mathcal{K}^t} \frac{P_{KL}^k(t,b)}{P_{KL}^{\mathcal{K}^t}(t,b)} \frac{RVA^k(t,b)}{RVA^{\mathcal{K}^t}(t,b)} \left(TFPROD_{VA}^k(t,b)\right)^{-1}, \tag{24}$$

or

$$TFPROD_{VA}^{\mathcal{K}^t}(t,b) = \left(\sum_{k \in \mathcal{K}^t} \frac{P_{KL}^k(t,b)}{P_{KL}^{\mathcal{K}^t}(t,b)} \frac{RVA^k(t,b)}{RVA^{\mathcal{K}^t}(t,b)} \left(TFPROD_{VA}^k(t,b)\right)^{-1} \right)^{-1}. \tag{25}$$

This is our alternative result. Thus, aggregate total factor productivity can also be obtained as a weighted *harmonic* sum of individual productivities, with weights

$$\psi^{kt} \equiv \frac{P_{KL}^k(t,b)}{P_{KL}^{\mathcal{K}^t}(t,b)} \frac{RVA^k(t,b)}{RVA^{\mathcal{K}^t}(t,b)} \quad (k \in \mathcal{K}^t). \tag{26}$$

Notice that these weights do not necessarily add up to 1. However, like above, the weights ψ^{kt} can be decomposed such that expression (25) can be considered as a weighted (harmonic) mean of adjusted individual productivities.

It is interesting to compare the structure of the two sets of weights ϕ^{kt} and ψ^{kt}. The former are based on real primary input shares and relative value-added price levels, whereas the latter are based on real output (value added) shares and relative primary input price levels.

Summarizing, there is no unique relation between the individual total factor productivities and the total factor productivity of the aggregate. One must either multiply the individual productivities by weights ϕ^{kt} and add up, or use weights ψ^{kt} and take the harmonic sum.

4.1 Additivity Imposed

We observed that both real valued added and real primary input are generally non-additive.[9] A sufficient condition for additivity is that deflators for the ensemble, for value added as well as primary input, are Paasche-type indices. This can be seen as

[9]Notice that we are considering here additivity of production units, which is different from additivity of commodities as considered in Balk (2016, Section 4.2).

follows. Additivity of real value added,

$$RVA^{\mathcal{K}^t}(t,b) = \sum_{k\in\mathcal{K}^t} RVA^k(t,b), \tag{27}$$

is, by inserting the definitions of real value added, equivalent to

$$\frac{1}{P_{VA}^{\mathcal{K}^t}(t,b)} = \sum_{k\in\mathcal{K}^t} \frac{VA^{kt}}{VA^{\mathcal{K}^t t}} \frac{1}{P_{VA}^{k}(t,b)}. \tag{28}$$

But this relation simply expresses that the value-added based deflator for the ensemble is a Paasche index of the deflators for the individual production units (recall that nominal value added is additive). Similarly, if the primary-input based deflator for the ensemble is a Paasche index of the unit-specific deflators, then

$$X_{KL}^{\mathcal{K}^t}(t,b) = \sum_{k\in\mathcal{K}^t} X_{KL}^{k}(t,b). \tag{29}$$

It is straightforward to check that if conditions (27) and (29) are satisfied, then instead of expression (18) we obtain the simpler expression[10]

$$TFPROD_{VA}^{\mathcal{K}^t}(t,b) = \sum_{k\in\mathcal{K}^t} \frac{X_{KL}^{k}(t,b)}{X_{KL}^{\mathcal{K}^t}(t,b)} TFPROD_{VA}^{k}(t,b), \tag{30}$$

and instead of expression (25) we obtain the simpler expression

$$TFPROD_{VA}^{\mathcal{K}^t}(t,b) = \left(\sum_{k\in\mathcal{K}^t} \frac{RVA^k(t,b)}{RVA^{\mathcal{K}^t}(t,b)} \left(TFPROD_{VA}^{k}(t,b)\right)^{-1} \right)^{-1}. \tag{31}$$

In both cases the weights now do add up to 1. The result is simple to summarize. If one weighs individual total factor productivities by real input shares then the arithmetic mean must be used, but if one weighs by real output shares then the harmonic mean must be used to arrive at an interpretable result.

Mixing this leads to unwanted effects. For example, combining the harmonic mean with real input shares leads to understating the productivity of the aggregate:

$$\left(\sum_{k\in\mathcal{K}^t} \frac{X_{KL}^{k}(t,b)}{X_{KL}^{\mathcal{K}^t}(t,b)} \left(TFPROD_{VA}^{k}(t,b)\right)^{-1} \right)^{-1} \leq TFPROD_{VA}^{\mathcal{K}^t}(t,b), \tag{32}$$

[10]This is the model underlying the CSLS decomposition as implemented by Calver and Murray (2016).

and combining the arithmetic mean with real output shares leads to overstating the productivity of the aggregate:

$$\sum_{k\in\mathcal{K}^t} \frac{RVA^k(t,b)}{RVA^{\mathcal{K}^t}(t,b)} TFPROD^k_{VA}(t,b) \geq TFPROD^{\mathcal{K}^t}_{VA}(t,b). \tag{33}$$

Both results rest on combining the mathematical fact that a harmonic mean is always less than or equal to an arithmetic mean with expressions (30) and (31). Equality in expressions (32) and (33) holds only when all the individual productivities $TFPROD^k_{VA}(t,b)$ $(k \in \mathcal{K}^t)$ are the same. Interestingly, the left-hand side of expression (33) is the target variable considered by Olley and Pakes (1996).

Also the type of mean matters. A geometric mean is greater than or equal to an harmonic mean, which implies that, using expression (31),

$$\prod_{k\in\mathcal{K}^t} \left(TFPROD^k_{VA}(t,b)\right)^{RVA^k(t,b)/RVA^{\mathcal{K}^t}(t,b)} \geq TFPROD^{\mathcal{K}^t}_{VA}(t,b). \tag{34}$$

Such a geometric mean was a target variable considered by Melitz and Polanec (2015). It is thus seen to overstate productivity of the aggregate.

Returning to the de Loecker and Konings (2006) case, it can be seen that instead of the right-hand side of expression (30) these authors considered

$$\sum_{k\in\mathcal{K}^t} \frac{L^{kt}}{L^{\mathcal{K}^t t}} TFPROD^k_{VA}(t,b), \tag{35}$$

which is a biased estimator of $TFPROD^{\mathcal{K}^t}_{VA}(t,b)$. The magnitude of the bias, and its sign, is of course an empirical matter.

5 Value-Added Based Labour Productivity

For value-added based labour productivity the setup of the previous section can simply be repeated. The only thing one needs to do is replacing real primary input by real labour input. Thus, for the individual production units real labour input is defined as

$$X^k_L(t,b) \equiv C^{kt}_L / P^k_L(t,b) \ (k \in \mathcal{K}^t). \tag{36}$$

Likewise, for the ensemble

$$X^{\mathcal{K}^t}_L(t,b) \equiv C^{\mathcal{K}^t t}_L / P^{\mathcal{K}^t}_L(t,b), \tag{37}$$

where $C^{\mathcal{K}^t t}_L \equiv \sum_{k\in\mathcal{K}^t} C^{kt}_L$ and $P^k_L(t,b)$ and $P^{\mathcal{K}^t}_L(t,b)$ are suitable deflators for the labour cost of the individual production units and the ensemble, respectively.

Labour productivity was defined as real value added divided by real labour input. Starting from the numerator of the labour productivity of the ensemble the decomposition appears to be

$$LPROD_{VA}^{\mathcal{K}^t}(t,b) = \sum_{k\in\mathcal{K}^t} \frac{P_{VA}^k(t,b)}{P_{VA}^{\mathcal{K}^t}(t,b)} \frac{X_L^k(t,b)}{X_L^{\mathcal{K}^t}(t,b)} LPROD_{VA}^k(t,b), \tag{38}$$

whereas starting from the denominator one obtains

$$LPROD_{VA}^{\mathcal{K}^t}(t,b) = \left(\sum_{k\in\mathcal{K}^t} \frac{P_L^k(t,b)}{P_L^{\mathcal{K}^t}(t,b)} \frac{RVA^k(t,b)}{RVA^{\mathcal{K}^t}(t,b)} \left(LPROD_{VA}^k(t,b)\right)^{-1} \right)^{-1}. \tag{39}$$

Both expressions relate value-added based labour productivity of the ensemble, considered as a higher level production unit, to the labour productivities of the constituent production units. Notice that the weights do not necessarily add up to 1. However, like shown before, these weights can be decomposed such that expressions (38) and (39) can be considered as weighted means of adjusted individual labour productivities.

5.1 Simple Labour Productivity

Two special cases deserve our attention. First, when for labour the simple sum quantity index is used then for the individual production units labour productivity is given by

$$LPROD_{VA}^k(t,b) = \frac{RVA^k(t,b)}{C_L^{kt}/P_L^k(t,b)} = \frac{RVA^k(t,b)}{C_L^{kb}Q_L^k(t,b)} = \frac{RVA^k(t,b)}{(C_L^{kb}/L^{kb})L^{kt}} \; (k \in \mathcal{K}^t), \tag{40}$$

and real labour input by $X_L^k(t,b) = (C_L^{kb}/L^{kb})L^{kt}$, where $L^{k\tau}$ denotes production unit k's total labour quantity at period τ ($\tau = b, t$). For the ensemble similar expressions hold.

Substitution, for the individual production units as well as for the ensemble, into expression (38) and some simplification delivers the following expression,

$$\frac{RVA^{\mathcal{K}^t}(t,b)}{L^{\mathcal{K}^t t}} = \sum_{k\in\mathcal{K}^t} \frac{P_{VA}^k(t,b)}{P_{VA}^{\mathcal{K}^t}(t,b)} \frac{L^{kt}}{L^{\mathcal{K}^t t}} \frac{RVA^k(t,b)}{L^{kt}}. \tag{41}$$

This is an expression in terms of simple labour productivities, as defined in Balk (2016). Put otherwise, expression (41) can be written as

$$SLPROD_{VA}^{\mathcal{K}^t}(t,b) = \sum_{k\in\mathcal{K}^t} \frac{P_{VA}^k(t,b)}{P_{VA}^{\mathcal{K}^t}(t,b)} \frac{L^{kt}}{L^{\mathcal{K}^t t}} SLPROD_{VA}^k(t,b). \tag{42}$$

It is quite natural to assume that the labour input of the ensemble, considered as a higher level production unit, is a simple sum of the labour inputs of the constituent units; that is, $L^{\mathcal{K}^t t} = \sum_{k \in \mathcal{K}^t} L^{kt}$. Then the fractions $L^{kt}/L^{\mathcal{K}^t t}$ $(k \in \mathcal{K}^t)$ are labour shares, adding up to 1. Notice, however, that these labour shares are premultiplied by relative price levels, so that the weights of the labour productivities themselves do not necessarily add up to 1. However, like shown before, these weights can be decomposed such that expression (42) can be considered as weighted mean of adjusted individual labour productivities. The relative price levels vanish when $P_{VA}^k(t,b) = P_{VA}^{\mathcal{K}^t}(t,b)$ $(k \in \mathcal{K}^t)$; that is, when there is no differential output price change among the production units.

There is, however, another way of looking at expression (42). To see this, recall that $(P_{VA}^k(t,b)/P_{VA}^{\mathcal{K}^t}(t,b))(RVA^k(t,b)/L^{kt})$ is so-called revenue labour productivity; that is, the result of deflating VA^{kt} not by its unit-k-specific deflator $P_{VA}^k(t,b)$ but by the ensemble-specific deflator $P_{VA}^{\mathcal{K}^t}(t,b)$. Weighing these revenue labour productivities by labour shares $L^{kt}/L^{\mathcal{K}^t t}$ then delivers aggregate labour productivity.

Finally, we notice that expression (41) is the model underlying the Generalized Exactly Additive Decomposition (GEAD) (see Balk 2016, Section 5.2). But it now turns out that an alternative decomposition can be developed.

To see this, notice that, by substituting expression (40) into expression (39) and using the product relation $C_L^{kt}/C_L^{kt'} = P_L^k(t,t')Q_L^k(t,t')$, expression (39) reduces to

$$SLPROD_{VA}^{\mathcal{K}^t}(t,b) = \left(\sum_{k \in \mathcal{K}^t} \frac{C_L^{kt}/L^{kt}}{C_L^{\mathcal{K}^t t}/L^{\mathcal{K}^t t}} \frac{RVA^k(t,b)}{RVA^{\mathcal{K}^t}(t,b)} \left(SLPROD_{VA}^k(t,b)\right)^{-1} \right)^{-1} . \quad (43)$$

This expression can be used to develop an alternative to the GEAD.

If the unit labour prices are the same across production units, that is, $C_L^{kt}/L^{kt} = \alpha$ $(k \in \mathcal{K}^t)$ and $C_L^{\mathcal{K}^t t}/L^{\mathcal{K}^t t} = \alpha$, then expression (43) further reduces to

$$SLPROD_{VA}^{\mathcal{K}^t}(t,b) = \left(\sum_{k \in \mathcal{K}^t} \frac{RVA^k(t,b)}{RVA^{\mathcal{K}^t}(t,b)} \left(SLPROD_{VA}^k(t,b)\right)^{-1} \right)^{-1} . \quad (44)$$

An alternative route to obtain this expression is the following. The assumption of equal unit labour prices across production units implies that $L^{\mathcal{K}^t t} = \sum_{k \in \mathcal{K}^t} L^{kt}$. Then, starting with this relation, dividing its left- and right-hand sides by $RVA^{\mathcal{K}^t}(t,b)$, and inserting at the right-hand side $RVA^k(t,b)/RVA^k(t,b) = 1$ $(k \in \mathcal{K}^t)$ one obtains expression (44).

Notice that the weights in expression (44) do not add up to 1, unless additivity holds. However, like shown before, these weights can be decomposed such that expression (44) can be considered as weighted (harmonic) mean of adjusted individual labour productivities.

5.2 *Additivity Imposed*

Second, let us assume that additivity holds; that is, $RVA^{\mathcal{K}^t}(t,b) = \sum_{k\in\mathcal{K}^t} RVA^k(t,b)$ and $L^{\mathcal{K}^t t} = \sum_{k\in\mathcal{K}^t} L^{kt}$. Instead of expression (38) we then obtain

$$SLPROD_{VA}^{\mathcal{K}^t}(t,b) = \sum_{k\in\mathcal{K}^t} \frac{L^{kt}}{L^{\mathcal{K}^t t}} SLPROD_{VA}^{k}(t,b), \tag{45}$$

and instead of expression (39) we obtain

$$SLPROD_{VA}^{\mathcal{K}^t}(t,b) = \left(\sum_{k\in\mathcal{K}^t} \frac{RVA^k(t,b)}{RVA^{\mathcal{K}^t}(t,b)} \left(SLPROD_{VA}^{k}(t,b)\right)^{-1} \right)^{-1}. \tag{46}$$

Now in both cases the weights add up to 1. The labour-share weighted arithmetic mean of simple labour productivities appears to be equal to the real-value-added-share weighted harmonic mean of simple labour productivities, and both are equal to the simple labour productivity of the aggregate.

Mixing means and weights leads to undesirable results. Using the general relation between harmonic and arithmetic means, we conclude that

$$\left(\sum_{k\in\mathcal{K}^t} \frac{L^{kt}}{L^{\mathcal{K}^t t}} \left(SLPROD_{VA}^{k}(t,b)\right)^{-1} \right)^{-1} \leq SLPROD_{VA}^{\mathcal{K}^t}(t,b) \tag{47}$$

$$\sum_{k\in\mathcal{K}^t} \frac{RVA^k(t,b)}{RVA^{\mathcal{K}^t}(t,b)} SLPROD_{VA}^{k}(t,b) \geq SLPROD_{VA}^{\mathcal{K}^t}(t,b). \tag{48}$$

Thus, a labour-share weighted harmonic mean of simple labour productivities understates labour productivity of the aggregate, while a real-value-added-share weighted arithmetic mean overstates this. The second inequality was also obtained by van Biesebroeck (2008), though in a less direct way.

Also here the type of mean matters. As a geometric mean is less then or equal to an arithmetic mean, we conclude that

$$\prod_{k\in\mathcal{K}^t} \left(SLPROD_{VA}^{k}(t,b)\right)^{L^{kt}/L^{\mathcal{K}^t t}} \leq SLPROD_{VA}^{\mathcal{K}^t}(t,b). \tag{49}$$

Such a geometric mean of simple labour productivities features prominently in Melitz and Polanec (2015). In the Appendix of their paper decompositions based on the left-hand side and the right-hand side of expression (49) are empirically compared. The geometric mean was also considered as target variable for firms by Hyytinen and Maliranta (2013) and Maliranta and Määttänen (2015).

Notice that the right-hand side of expression (45) is the target variable of the TRAD and CSLS decompositions considered in Balk (2016, Section 5.2).

Thus these decompositions are consistent; that is, aggregate productivity can be interpreted as productivity of the aggregate. However, underlying this result is the assumption of additivity, which is quite restrictive.

6 Gross-Output Based Productivity

There are not so many microdata studies dealing with the concept of gross-output based productivity. For any individual production unit gross-output based total factor productivity is defined as

$$TFPROD_Y^k(t,b) \equiv \frac{Y^k(t,b)}{X_{KLEMS}^k(t,b)} = \frac{R^{kt}/P_R^k(t,b)}{C_{KLEMS}^{kt}/P_{KLEMS}^k(t,b)} \; (k \in \mathcal{K}^t). \tag{50}$$

In the numerator we have real revenue $Y^k(t,b)$; that is, nominal revenue R^{kt} deflated by a k-specific revenue based price index with reference period b, $P_R^k(t,b)$. In the denominator we have real KLEMS input $X_{KLEMS}^k(t,b)$; that is, nominal KLEMS input cost C_{KLEMS}^{kt} deflated by a k-specific KLEMS input based price index with the same reference period, $P_{KLEMS}^k(t,b)$; so that the ratio $TFPROD_Y^k(t,b)$ is a dimensionless variable.

Similarly, gross-output based simple labour productivity is defined as

$$SLPROD_Y^k(t,b) \equiv \frac{Y^k(t,b)}{L^{kt}} \; (k \in \mathcal{K}^t); \tag{51}$$

that is, real revenue per unit of labour. The dimension of this variable is money of reference period b.

Suppose that we have access to production-unit specific data such that either of these measures can be compiled. Which weights would be appropriate? We review a number of typical studies.

6.1 Simple Labour Productivity

Let us start with the target variable considered by Baily et al. (BBH) (2001). This is $SLPROD_Y^k(t,b)$, though instead of unit-specific deflators industry-level deflators were used. The labour unit was an hour worked. These simple labour productivities were weighed by *labour* shares; that is, by $L^{kt}/L^{\mathcal{K}^t t} = L^{kt}/\sum_{k\in\mathcal{K}^t} L^{kt}$. Thus, aggregate productivity was compiled as[11]

[11] This measure was also considered by Foster et al. (2001). Actually, two variants were considered, one where the labour unit is an hour worked and one where it is a worker. The geometric alternative was employed by Hyytinen and Maliranta (2013) for plants; labour quantity was thereby measured in full time equivalents.

$$LPROD_{BBH}^{\mathcal{K}^t}(t,b) \equiv \sum_{k\in\mathcal{K}^t} \frac{L^{kt}}{\sum_{k\in\mathcal{K}^t} L^{kt}} SLPROD_Y^k(t,b) = \frac{\sum_{k\in\mathcal{K}^t} Y^k(t,b)}{L^{\mathcal{K}^t t}}. \tag{52}$$

But what precisely does this mean? To see this, we must return to the accounting identities discussed in Sect. 2 and notice that

$$\sum_{k\in\mathcal{K}^t} R^{kt} = \sum_{k\in\mathcal{K}^t} \sum_{k'\in\mathcal{K}^t, k'\neq k} R^{kk't} + R^{\mathcal{K}^t t}. \tag{53}$$

Thus, total revenue is the sum of revenue obtained by internal deliveries (recall that $R^{kk't}$ is the revenue obtained by unit k from delivering to unit k') and aggregate revenue $R^{\mathcal{K}^t t}$, which is the revenue obtained by the ensemble $\mathcal{K}^t$, when the ensemble is considered as a consolidated production unit. Now, imposing additivity, that is, defining the aggregate revenue-based price index as a Paasche index of the k-specific revenue based price indices,

$$\frac{1}{P_R^{\mathcal{K}^t}(t,b)} \equiv \sum_{k\in\mathcal{K}^t} \frac{R^{kt}}{\sum_{k\in\mathcal{K}^t} R^{kt}} \frac{1}{P_R^k(t,b)}, \tag{54}$$

implies that expression (53) can be written as

$$P_R^{\mathcal{K}^t}(t,b) \sum_{k\in\mathcal{K}^t} Y^k(t,b) = \sum_{k\in\mathcal{K}^t} \sum_{k'\in\mathcal{K}^t, k'\neq k} R^{kk't} + R^{\mathcal{K}^t t}, \tag{55}$$

or

$$\sum_{k\in\mathcal{K}^t} Y^k(t,b) = \frac{\sum_{k\in\mathcal{K}^t} \sum_{k'\in\mathcal{K}^t, k'\neq k} R^{kk't}}{P_R^{\mathcal{K}^t}(t,b)} + \frac{R^{\mathcal{K}^t t}}{P_R^{\mathcal{K}^t}(t,b)}. \tag{56}$$

If we define real revenue of the ensemble $\mathcal{K}^t$, considered as a consolidated production unit, by $Y^{\mathcal{K}^t}(t,b) \equiv R^{\mathcal{K}^t t}/P_R^{\mathcal{K}^t}(t,b)$, then expression (56) can be simplified to

$$\sum_{k\in\mathcal{K}^t} Y^k(t,b) = \frac{\sum_{k\in\mathcal{K}^t} \sum_{k'\in\mathcal{K}^t, k'\neq k} R^{kk't}}{P_R^{\mathcal{K}^t}(t,b)} + Y^{\mathcal{K}^t}(t,b). \tag{57}$$

Substituting expression (57) into expression (52) and applying definition (51) to the ensemble considered as a production unit delivers the following relation:

$$LPROD_{BBH}^{\mathcal{K}^t}(t,b) = SLPROD_Y^{\mathcal{K}^t}(t,b) \left(1 + \frac{\sum_{k\in\mathcal{K}^t} \sum_{k'\in\mathcal{K}^t, k'\neq k} R^{kk't}}{R^{\mathcal{K}^t t}}\right). \tag{58}$$

Since nominal revenue is non-negative, it appears that aggregate BBH productivity overstates simple labour productivity of the aggregate, and that the magnitude of the bias depends on the relative extent of the intra-ensemble deliveries. The bias vanishes only when there are no intra-ensemble deliveries.

Foster et al. (FHK) (2001) considered simple labour productivities weighed by *real output* shares; that is,

$$LPROD_{FHK}^{\mathcal{K}^t}(t,b) \equiv \sum_{k\in\mathcal{K}^t} \frac{Y^k(t,b)}{\sum_{k\in\mathcal{K}^t} Y^k(t,b)} SLPROD_Y^k(t,b). \tag{59}$$

Applying the arithmetic-harmonic mean inequality, and definitions (51) and (52) respectively, we obtain

$$LPROD_{FHK}^{\mathcal{K}^t}(t,b) \geq \frac{\sum_{k\in\mathcal{K}^t} Y^k(t,b)}{L^{\mathcal{K}^t t}} = LPROD_{BBH}^{\mathcal{K}^t}(t,b). \tag{60}$$

The right-hand side is familiar from the foregoing. Combining expressions (60) and (58) we may conclude that, even in the case of industries exhibiting no intra-ensemble trade, $LPROD_{FHK}^{\mathcal{K}^t}(t,b)$ overstates simple labour productivity of the aggregate, $SLPROD_Y^{\mathcal{K}^t}(t,b)$.

6.2 Total Factor Productivity

We now turn to $TFPROD_Y^k(t,b)$, a key variable considered by Bartelsman and Dhrymes (BD) (1998). They had industry and time effects removed econometrically, but that does not need to concern us here. The individual gross-output based total factor productivities were weighed by *real KLEMS input* shares $X_{KLEMS}^k(t,b)/\sum_{k\in\mathcal{K}^t} X_{KLEMS}^k(t,b)$, so that aggregate total factor productivity was compiled as

$$\begin{aligned} TFPROD_{BD}^{\mathcal{K}^t}(t,b) &\equiv \sum_{k\in\mathcal{K}^t} \frac{X_{KLEMS}^k(t,b)}{\sum_{k\in\mathcal{K}^t} X_{KLEMS}^k(t,b)} TFPROD_Y^k(t,b) \\ &= \frac{\sum_{k\in\mathcal{K}^t} Y^k(t,b)}{\sum_{k\in\mathcal{K}^t} X_{KLEMS}^k(t,b)}. \end{aligned} \tag{61}$$

Notice that, assuming that additivity at the output side holds, the numerator is given by expression (57). For the denominator a similar expression can be derived. To see this, we again return to the accounting identities in Sect. 2 and notice that

$$\sum_{k\in\mathcal{K}^t} C_{EMS}^{kt} = \sum_{k\in\mathcal{K}^t} \sum_{k'\in\mathcal{K}^t, k'\neq k} C_{EMS}^{k'kt} + C_{EMS}^{\mathcal{K}^t t}, \tag{62}$$

Adding at both sides $C_{KL}^{\mathcal{K}^t t} = \sum_{k\in\mathcal{K}^t} C_{KL}^{kt}$, we obtain the following accounting relation:

$$\sum_{k\in\mathcal{K}^t} C_{KLEMS}^{kt} = \sum_{k\in\mathcal{K}^t} \sum_{k'\in\mathcal{K}^t, k'\neq k} C_{EMS}^{k'kt} + C_{KLEMS}^{\mathcal{K}^t t}. \tag{63}$$

Thus, total cost is the sum of cost incurred by internal deliveries (recall that $C_{EMS}^{k'kt}$ is the cost incurred by unit k for purchases from unit k') and aggregate cost $C_{KLEMS}^{\mathcal{K}'t}$, which is the KLEMS input cost of the ensemble $\mathcal{K}^t$, considered as a consolidated production unit. Now, imposing additivity at the input side, that is, defining the aggregate KLEMS input based price index as a Paasche index of the k-specific KLEMS input based price indices,

$$\frac{1}{P_{KLEMS}^{\mathcal{K}^t}(t,b)} \equiv \sum_{k\in\mathcal{K}^t} \frac{C_{KLEMS}^{kt}}{\sum_{k\in\mathcal{K}^t} C_{KLEMS}^{kt}} \frac{1}{P_{KLEMS}^{k}(t,b)}, \tag{64}$$

implies that expression (63) can be written as

$$P_{KLEMS}^{\mathcal{K}^t}(t,b) \sum_{k\in\mathcal{K}^t} X_{KLEMS}^{k}(t,b) = \sum_{k\in\mathcal{K}^t} \sum_{k'\in\mathcal{K}^t, k'\neq k} C_{EMS}^{k'kt} + C_{KLEMS}^{\mathcal{K}'t}, \tag{65}$$

or

$$\sum_{k\in\mathcal{K}^t} X_{KLEMS}^{k}(t,b) = \frac{\sum_{k\in\mathcal{K}^t} \sum_{k'\in\mathcal{K}^t, k'\neq k} C_{EMS}^{k'kt}}{P_{KLEMS}^{\mathcal{K}^t}(t,b)} + \frac{C_{KLEMS}^{\mathcal{K}'t}}{P_{KLEMS}^{\mathcal{K}^t}(t,b)}. \tag{66}$$

If we define real KLEMS input of the ensemble $\mathcal{K}^t$, considered as a consolidated production unit, as $X_{KLEMS}^{\mathcal{K}^t}(t,b) \equiv C_{KLEMS}^{\mathcal{K}'t}/P_{KLEMS}^{\mathcal{K}^t}(t,b)$, then expression (66) can be simplified to

$$\sum_{k\in\mathcal{K}^t} X_{KLEMS}^{k}(t,b) = \frac{\sum_{k\in\mathcal{K}^t} \sum_{k'\in\mathcal{K}^t, k'\neq k} C_{EMS}^{k'kt}}{P_{KLEMS}^{\mathcal{K}^t}(t,b)} + X_{KLEMS}^{\mathcal{K}^t}(t,b). \tag{67}$$

Substituting expressions (57) and (67) into expression (61) and applying definition (50) to the ensemble considered as a production unit delivers the following relation:

$$\begin{aligned} &TFPROD_{BD}^{\mathcal{K}^t}(t,b) \\ &= TFPROD_{Y}^{\mathcal{K}^t}(t,b) \frac{1 + \sum_{k\in\mathcal{K}^t} \sum_{k'\in\mathcal{K}^t, k'\neq k} R^{kk't}/R^{\mathcal{K}'t}}{1 + \sum_{k\in\mathcal{K}^t} \sum_{k'\in\mathcal{K}^t, k'\neq k} C_{EMS}^{k'kt}/C_{KLEMS}^{\mathcal{K}'t}}. \end{aligned} \tag{68}$$

As observed in Sect. 2, National Accounting conventions imply that revenue and cost of the intra-ensemble transactions are equal; that is,

$$\sum_{k\in\mathcal{K}^t} \sum_{k'\in\mathcal{K}^t, k'\neq k} R^{kk't} = \sum_{k\in\mathcal{K}^t} \sum_{k'\in\mathcal{K}^t, k'\neq k} C_{EMS}^{k'kt}.$$

Thus the magnitude of the bias of aggregate BD total factor productivity depends on the magnitude of aggregate revenue $R^{\mathcal{K}'t}$ relative to aggregate KLEMS input cost $C_{KLEMS}^{\mathcal{K}'t}$. Put otherwise, the magnitude of the bias depends on aggregate profit $\Pi^{\mathcal{K}'t}$.

If aggregate profit is positive (negative), then aggregate BD total factor productivity understates (overstates) total factor productivity of the aggregate. If aggregate profit equals 0, then the bias vanishes. A sufficient condition for zero aggregate profit is that $\Pi^{kt} = 0$ for each individual production unit $k \in \mathcal{K}^t$. Of course, the bias also vanishes in the trivial case when there are no intra-ensemble deliveries.

Foster et al. (2001) considered total factor productivities weighed by *real output* shares; that is,[12]

$$TFPROD^{\mathcal{K}^t}_{FHK}(t,b) \equiv \sum_{k \in \mathcal{K}^t} \frac{Y^k(t,b)}{\sum_{k \in \mathcal{K}^t} Y^k(t,b)} TFPROD^k_Y(t,b). \tag{69}$$

Applying the arithmetic-harmonic mean inequality and using the definitions in expressions (50) and (61), we find that

$$TFPROD^{\mathcal{K}^t}_{FHK}(t,b) \geq TFPROD^{\mathcal{K}^t}_{BD}(t,b). \tag{70}$$

Now expression (68) above tells us that, under additivity at the input and the output side, $TFPROD^{\mathcal{K}^t}_{BD}(t,b)$ is an unbiased measure of total factor productivity of the aggregate if there are no intra-ensemble deliveries. Thus, we may conclude that in the cases studied by Foster et al., which were four-digit level industries where intra-industry deliveries are unlikely, $TFPROD^{\mathcal{K}^t}_{FHK}(t,b)$ most likely overstates total factor productivity of the aggregate, $TFPROD^{\mathcal{K}^t}_{Y}(t,b)$.

The target variable of Eslava et al. (EHKK) (2013) appears to be

$$TFPROD^{\mathcal{K}^t}_{EHKK}(t,b) \equiv \prod_{k \in \mathcal{K}^t} (TFPROD^k_Y(t,b))^{Y^k(t,b)/\sum_{k \in \mathcal{K}^t} Y^k(t,b)}; \tag{71}$$

that is, the geometric variant of the FHK measure defined by expression (69). Using subsequently the geometric-harmonic mean inequality, definition (50), and expression (61), we obtain

$$TFPROD^{\mathcal{K}^t}_{EHKK}(t,b) \geq TFPROD^{\mathcal{K}^t}_{BD}(t,b). \tag{72}$$

As we have seen, the right-hand side of this expression may or may not approximate $TFPROD^{\mathcal{K}^t}_{Y}(t,b)$.

It is now interesting to consider a recent paper by Collard-Wexler and de Loecker (CWL) (2015). These authors also dealt with $TFPROD^k_Y(t,b)$ ($k \in \mathcal{K}^t$), but to obtain aggregate productivity the individual total factor productivities were weighed by *nominal revenue* shares $R^{kt}/\sum_{k \in \mathcal{K}^t} R^{kt}$. Thus aggregate productivity was defined as

$$TFPROD^{\mathcal{K}^t}_{CWL}(t,b) \equiv \sum_{k \in \mathcal{K}^t} \frac{R^{kt}}{\sum_{k \in \mathcal{K}^t} R^{kt}} TFPROD^k_Y(t,b). \tag{73}$$

[12] Actually, their multi-factor productivity index, discussed in the extended version of this paper, can be seen as a special case of $TFPROD^k_Y(t,b)$.

To obtain an interpretation for this mean, we first relate it to the alternative where real shares $Y^k(t,b)/\sum_{k\in\mathcal{K}^t} Y^k(t,b)$ are used as weights,

$$TFPROD^{\mathcal{K}^t}_{CWL}(t,b) = \sum_{k\in\mathcal{K}^t} \frac{Y^k(t,b)}{\sum_{k\in\mathcal{K}^t} Y^k(t,b)} TFPROD^k_Y(t,b) + \sum_{k\in\mathcal{K}^t} \left(\frac{R^{kt}}{\sum_{k\in\mathcal{K}^t} R^{kt}} - \frac{Y^k(t,b)}{\sum_{k\in\mathcal{K}^t} Y^k(t,b)} \right) TFPROD^k_Y(t,b). \quad (74)$$

The first term at the right-hand side of this equation is familiar; it is the FHK measure as defined by expression (69). The second term has the form of a covariance, but there is in general no compelling reason for this covariance to be positive or negative, large or small. Taken together, on the assumption that the covariance in Eq. (74) equals 0, it seems likely that aggregate CWL productivity overstates the productivity of the aggregate.

6.3 Some Empirical Comparisons

The primary purpose of the classic paper by Foster et al. (2001) was to compare decompositions of intertemporal change of the three aggregate measures $TFPROD^{\mathcal{K}^t}_{FHK}(t,b)$, $LPROD^{\mathcal{K}^t}_{FHK}(t,b)$, and $LPROD^{\mathcal{K}^t}_{BBH}(t,b)$. They specifically examined the Foster-Haltiwanger-Krizan (FHK) and the Griliches-Regev (GR) decomposition methods (see Balk (2016), expressions (2.43) and (2.50) respectively). It turned out that, though the levels were of course different, the FHK decompositions of $\Delta TFPROD^{\mathcal{K}^t}_{FHK}(t,b)$ and $\Delta LPROD^{\mathcal{K}^t}_{FHK}(t,b)$ were strikingly similar. The levels as well as the FHK decompositions of $\Delta LPROD^{\mathcal{K}^t}_{FHK}(t,b)$ and $\Delta LPROD^{\mathcal{K}^t}_{BBH}(t,b)$ differed, however, remarkably. Interestingly, for the three aggregate measures the GR decomposition delivered almost the same results. Overall, the 'within' term appeared dominant.

7 Conclusion

Our overall conclusion is that not every combination of micro-, or meso-level productivities, weights, and aggregator function (mean) leads to a nice interpretation of aggregate productivity as productivity of the aggregate. Specifically:

- An arithmetic 'mean' of value-added based total factor productivities requires weights based on relative real primary input times relative value-added based price levels.

- A harmonic 'mean' of value-added based total factor productivities requires weights based on relative real value added times relative primary input price levels.
- Under additivity the relative price levels disappear from the expressions.
- Similar results hold for value-added based (simple) labour productivities.
- An arithmetic mean of gross-output (revenue) based simple labour productivities weighed with (physical) labour input shares is likely to overstate its aggregate counterpart.
- An arithmetic mean of gross-output (revenue) based total factor productivities weighed with real input shares approximates gross-output (revenue) based total factor productivity of the aggregate; the magnitude of the bias depends on aggregate profit.

References

Baily, M. N., Bartelsman, E. J., & Haltiwanger, J. (2001). Labor productivity: Structural change and cyclical dynamics. *The Review of Economics and Statistics, 83*, 420–433.

Balk, B. M. (2010). An assumption-free framework for measuring productivity change. *The Review of Income and Wealth, 56*(Special Issue 1), S224–S256.

Balk, B. M. (2011). Measuring and decomposing capital input cost. *The Review of Income and Wealth, 57*, 490–512.

Balk, B. M. (2014). Dissecting aggregate output and labour productivity change. *Journal of Productivity Analysis, 42*, 35–43.

Balk, B. M. (2015). Measuring and relating aggregate and subaggregate total factor productivity change without neoclassical assumptions. *Statistica Neerlandica, 69*, 21–28.

Balk, B. M. (2016). The dynamics of productivity change: A review of the bottom-up approach. In W. H. Greene, L. Khalaf, R. C. Sickles, M. Veall, & M.-C. Voia (Eds.), *Productivity and efficiency analysis*. Proceedings in Business and Economics. Cham: Springer International Publishing.

Bartelsman, E. J., & Dhrymes, Ph. J. (1998). Productivity dynamics: US manufacturing plants, 1972–1986. *Journal of Productivity Analysis, 9*, 5–34.

Basu, S., & Fernald, J. G. (2002). Aggregate productivity and aggregate technology. *European Economic Review, 46*, 963–991.

Calver, M., & Murray, A. (2016). Decomposing multifactor productivity growth in Canada by industry and province, 1997–2014. *International Productivity Monitor 31*, 88–112.

Collard-Wexler, A., & de Loecker, J. (2015). Reallocation and technology: Evidence from the U. S. steel industry. *American Economic Review, 105*, 131–171.

de Loecker, J., & Konings, J. (2006). Job reallocation and productivity growth in a post-socialist economy: Evidence from Slovenian manufacturing. *European Journal of Political Economy, 22*, 388–408.

Diewert, W. E. (1980). Aggregation problems in the measurement of capital. In D. Usher (Ed.), *The measurement of capital*. Cambridge, MA: National Bureau of Economic Research/University of Chicago Press.

Dumagan, J. C., & Balk, B. M. (2016). Dissecting aggregate output and labour productivity change: A postscript on the role of relative prices. *Journal of Productivity Analysis, 45*, 117–119.

Eslava, M., Haltiwanger, J., Kugler, A., & Kugler, M. (2013). Trade and market selection: Evidence from manufacturing plants in Colombia. *Review of Economic Dynamics, 16*, 135–158.

Foster, L., Haltiwanger, J., & Krizan, C. J. (2001). Aggregate productivity growth: Lessons from microeconomic evidence. In C. R. Hulten, E. R. Dean, & M. J. Harper (Eds.), *New developments in productivity analysis* (Studies in Income and Wealth, Vol. 63). Chicago/London: The University of Chicago Press.

Hyytinen, A., & Maliranta, M. (2013). Firm lifecycles and evolution of industry productivity. *Research Policy, 42*, 1080–1098.

Maliranta, M., & Määttänen, N. (2015). An augmented static Olley-Pakes productivity decomposition with entry and exit: Measurement and interpretation. *Economica, 82*, 1372–1416.

Melitz, M. J., & Polanec, S. (2015). Dynamic Olley-Pakes productivity decomposition with entry and exit. *Rand Journal of Economics, 46*, 362–375.

Olley, S., & Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica, 64*, 1263–1297.

PPI Manual. (2004). *Producer price index manual: Theory and practice* (Published for ILO, IMF, OECD, UN, Eurostat, The World Bank by Washington, DC: IMF).

van Biesebroeck, J. (2008). Aggregating and decomposing productivity. *Review of Business and Economics, LIII*, 122–146.

Vancauteren, M., Veldhuizen, E., & Balk, B. M. (2012). *Measures of productivity change: Which outcome do you want?* Paper presented at the 32nd General Conference of the IARIW, Boston, MA, August 5–11, 2012.

Confidence Sets for Inequality Measures: Fieller-Type Methods

Jean-Marie Dufour, Emmanuel Flachaire, Lynda Khalaf, and Abdallah Zalghout

Abstract Asymptotic and bootstrap inference methods for inequality indices are for the most part unreliable due to the complex empirical features of the underlying distributions. In this paper, we introduce a Fieller-type method for the Theil Index and assess its finite-sample properties by a Monte Carlo simulation study. The fact that almost all inequality indices can be written as a ratio of functions of moments and that a Fieller-type method does not suffer from weak identification as the denominator approaches zero, makes it an appealing alternative to the

This work was supported by the William Dow Chair of Political Economy (McGill University), the Bank of Canada (Research Fellowship), the Toulouse School of Economics (Pierre-de-Fermat Chair of excellence), the Universitad Carlos III de Madrid (Banco Santander de Madrid Chair of excellence), the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, the Fonds de recherche sur la société et la culture (Québec), and by project ANR-16-CE41-0005 managed by the French National Research Agency (ANR).

J.-M. Dufour
Department of Economics, McGill University, Centre interuniversitaire de recherche en analyse des organisations (CIRANO), and Centre interuniversitaire de recherche en économie quantitative (CIREQ), Montréal, QC, Canada
e-mail: jean-marie.dufour@mcgill.ca

E. Flachaire
CNRS, EHESS, Centrale Marseille, AMSE, Aix-Marseille Université, Marseille, France
e-mail: emmanuel.flachaire@univ-amu.fr

L. Khalaf (✉)
Department of Economics, Centre interuniversitaire de recherche en économie quantitative (CIREQ), Carleton University, Ottawa, ON, Canada

Groupe de recherche en économie de l'énergie, de l'environnement et des ressources naturelles (GREEN), Université Laval, Quebec City, QC, Canada
e-mail: Lynda_Khalaf@carleton.ca

A. Zalghout
Economics Department, Carleton University, Ottawa, ON, Canada
e-mail: abdullahzalghout@cmail.carleton.ca

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_6

available inference methods. Our simulation results exhibit several cases where a Fieller-type method improves coverage. This occurs in particular when the Data Generating Process (DGP) follows a finite mixture of distributions, which reflects irregularities arising from low observations (close to zero) as opposed to large (right-tail) observations. Designs that forgo the interconnected effects of both boundaries provide possibly misleading finite-sample evidence. This suggests a useful prescription for simulation studies in this literature.

Keywords Inequality measures · Fieller-type confidence set · Delta method · Singh-Maddala distribution · Gamma distribution · Mixture

1 Introduction

Asymptotic inference methods for inequality indices are for the most part unreliable due to the complex empirical features of the underlying distributions, particularly in the case of income. Typically, the presence of heavy tails invalidates standard parametric and nonparametric inference methods based on central limit theory (CLT), leading to spurious conclusions with samples of realistic size. This problem persists even with very large samples. Moreover, for some parameter values, the moments of widely used distributions in this literature, such as the Singh-Maddala and Pareto distributions, do not exist. Early references can be traced back to Maasoumi (1997) or Mills and Zandvakili (1997); for a survey, see Cowell and Flachaire (2015).

Bootstrap inference methods emerge as an appealing alternative, since observations can often be viewed as independent random draws from the population. The first study to use and recommend bootstrap methods for inequality indices is the one of Mills and Zandvakili (1997). Biewen (2002) studied the performance of standard bootstrap methods in the context of inequality measures assuming a lognormal distribution as the Data Generating Process (DGP). Although his results suggest that the bootstrap performs well in finite samples, the lognormal distribution he used does not capture the thick tails typically observed in empirical work (Davidson and Flachaire 2007). Other simulation studies based on heavy-tailed distributions, such as the Singh-Maddala distribution, confirm that bootstrapping fails – often by far – to control coverage rates, despite the fact that they lead to higher-order refinements relative to asymptotic methods (Davidson and Flachaire 2007; Cowell and Flachaire 2007).

Non-standard inference methods have recently been suggested in an attempt to improve the quality of inference for inequality measures. Two notable approaches are permutation tests (Dufour et al. 2017) and semi-parametric methods (Davidson and Flachaire 2007; Cowell and Flachaire 2007). The permutational approach focuses on testing the equality of two indices, and the authors show that it performs very well when the two indices come from similar distributions. The semi-parametric bootstrap approach assumes a parametric distribution for the right tail

and a nonparametric empirical distribution function (EDF) for the rest. This method leads to considerable refinement over their asymptotic and bootstrap counterparts, provided the probability of the tail (p) and the ordered statistics defining the upper tail (k) are well chosen, which is usually not an easy task. Thus except for very specific cases, accurate inference methods on inequality measures are not available.

In this paper, we introduce the Fieller method for the Theil Index, and we assess its finite-sample properties through a Monte Carlo simulation study. Fieller's method was originally introduced for inference on the ratio of two means of normal variates. It is based on inverting a t-test of a linear restriction associated with the ratio, and allows one to get exact confidence sets for this ratio. This holds promise relative to the standard Delta method especially when the denominator of the ratio approaches zero, since the implicit linear reformulation addresses the underlying weak identification. Most inequality indices can be written as a ratio of functions of moments; so a Fieller-type method may plausibly lead to more reliable inference on these indices. However, given the non-linear dependence between the numerator and the denominator of the indices along with the typically positive support of the underlying distributions, the advantages from employing a Fieller-type method should not be taken for granted. This motivates the present work.

The method first introduced by Fieller (1940, 1954) was extended to independent samples of different sizes (Bennett 1953), multivariate models (Bennett 1959; Zerbe et al. 1982), general exponential regression models (Cox 1967), general linear regression models (Zerbe 1978; Dufour 1997), and dynamic models with possibly persistent covariates (Bernard et al. 2007, 2015; Stock and Lazarus 2016). Bolduc et al. (2010) used several variants of Fieller's approach to build simultaneous confidence sets for multiple (possibly weakly identified) ratios and they showed in a simulation study that a Fieller-type method outperforms the Delta method and controls level globally. Empirically, Fieller's approach has been routinely applied in medical research and to a lesser extent in economics (Srivastava 1986; Willan and O'Brien 1996; Johannesson et al. 1996; Laska et al. 1997; Stock and Lazarus 2016).

Fieller-type confidence sets may be perceived as counter-intuitive, because they can produce unbounded regions including the whole real line.[1] This perhaps gives reason for their unpopularity in applied work relative to Delta method-based confidence sets (DCS), despite their solid theoretical foundation. However, the geometric interpretation of Fieller's method is quite intuitive (see von Luxburg and Franz 2004). More to the point here, in the presence of identification problems, valid coverage requires possibly unbounded outcomes (Gleser and Hwang 1987; Dufour 1997), which is allowed by a Fieller-type solution as opposed to the Delta method.

Our simulation results provide evidence on the superiority of a Fieller-type method in terms of reducing size distortions in many useful cases. In particular, a Fieller-type method improves coverage over the Delta method when the distribution under the null allows for bunching of low observations (close to zero) in addition to a thick right tail. For such cases, the denominator of the Theil index is small relative

[1] See Scheffé (1970) for a modified version of Fieller's method that avoid the confidence set $\mathbb{R}$.

to the numerator and inequality is high. Methodologically, our findings suggest that studies focusing only on the upper tail may misrepresent finite-sample distortions with positive support distributions. In contrast, our design does allow us to assess further irregularities arising from low observations. As illustrated by Cowell and Victoria-Feser (1996), both boundaries may matter for general entropy class of indices, although not necessarily for the Theil index.

The paper is organized as follows. Section 2 presents the Fieller-type method for the Theil index. In Sect. 3, Monte Carlo results are provided. Section 4 concludes.

2 Fieller-Type Inference for Inequality Measures

Most income inequality indices depend solely on the underlying distribution of income. Technically speaking, they can be typically written as a functional which maps the space of cumulative distribution functions (CDFs) of income to the positive real line.

2.1 General Functional Ratios

Denote by Y the random variable representing income, and by $F_Y(y)$ its CDF. The class of indices considered in this paper can be written as the ratio of functions of two moments, namely the mean μ and another moment $\nu = \mathbb{E}[\phi(Y)]$, where $\phi(\cdot)$ is a given function. In particular, for the Theil index $\phi(Y) = Y \log(Y)$. In general, most inequality indices can be written as

$$I = \psi(\mu;\, \nu) = \frac{\psi_1(\mu;\, \nu)}{\psi_2(\mu;\, \nu)}. \tag{1}$$

The index I can be estimated using sample moments:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i\,, \quad \hat{\nu} = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i)\,, \tag{2}$$

where $Y_1, \ldots, Y_n$ is a sample of observations on Y, and $\phi(Y_i)$ is a function that takes different forms for different inequality indices. If we assume that the estimator is asymptotically normal, then the asymptotic covariance matrix can be estimated by

$$V(\hat{I}) = \frac{1}{n}\begin{bmatrix} \frac{\partial\psi}{\partial\mu} & \frac{\partial\psi}{\partial\nu} \end{bmatrix} \begin{bmatrix} \hat{\sigma}_\nu^2 & \hat{\sigma}_{\mu\nu} \\ \hat{\sigma}_{\mu\nu} & \hat{\sigma}_\nu^2 \end{bmatrix} \begin{bmatrix} \frac{\partial\psi}{\partial\mu} \\ \frac{\partial\psi}{\partial\nu} \end{bmatrix} \tag{3}$$

where $\hat{V}(\hat{I}) \equiv V(\hat{I})|_{\mu=\hat{\mu};\nu=\hat{\nu}}$. Here, $\hat{\sigma}_{\mu}^2$, $\hat{\sigma}_{\nu}^2$ and $\hat{\sigma}_{\mu\nu}$ are, respectively, estimates of the variance of Y, the variance of $\phi(Y)$, and the covariance of Y and $\phi(Y)$.[2]

In this paper, we consider the problem of building Fieller-type confidence sets (FCS) and Delta method confidence sets (DCS) for an index of the form I in (1). In general, this can be viewed as equivalent to finding the values of I_0 which are not rejected when one tests null hypotheses of the form

$$H_0(I_0) : \frac{\psi_1(\mu,\ \nu)}{\psi_2(\mu,\ \nu)} = I_0 \tag{4}$$

where I_0 is any admissible value of I. Here, this can be achieved by inverting the absolute value or the square of the relevant t-type statistic. To invert a t-test with respect to the parameter tested, we collect all the values of this parameter for which the test is not significant at a given level.

Following the Delta method, we invert the test statistic

$$t(I_0)^2 = \frac{(\hat{I} - I_0)^2}{\hat{V}(\hat{I})} \tag{5}$$

which leads to the confidence set

$$\text{DCS}(I; 1-\alpha) = \left[\hat{I} - z_{\alpha/2}[\hat{V}(\hat{I})]^{1/2}\,;\, \hat{I} + z_{\alpha/2}[\hat{V}(\hat{I})^{1/2}]\right] \tag{6}$$

where $z_{\alpha/2}$ is the usual α critical point based on the normal distribution (i.e., $\mathbb{P}[Z \geq z_{\alpha/2}] = \alpha/2$ for $Z \sim N[0,\ 1]$).

By contrast, the Fieller approach can be applied as follows. For each possible value I_0, the Fieller-type approach consists in considering the equivalent linear hypothesis

$$H_L(I_0) : \theta(I_0) = 0\,, \quad \text{where } \theta(I_0) = \psi_1(\mu,\ \nu) - I_0\,\psi_2(\mu,\ \nu) \tag{7}$$

where the superscript L is added to differentiate the original null hypothesis from its linear reformulation. Through this exact linearization, the Fieller-type method avoids possible (weak) identification problems when the denominator $\psi_2(\mu, \nu)$ is close to zero. To construct the FCS, we consider the square of the t-statistic associated with $H_L(I_0)$ in (7):

$$t(I_0)^2 = \frac{\hat{\theta}(I_0)^2}{\hat{V}[\hat{\theta}(I_0)]} \tag{8}$$

[2]Note that the variance of $\hat{\mu}$, the variance of $\hat{\nu}$ and covariance of the $(\hat{\mu},\ \hat{\nu})$ are equal to $\hat{\sigma}_{\mu}^2/n$, $\hat{\sigma}_{\nu}^2/n$ and $\hat{\sigma}_{\mu\nu}/n$.

where $\hat{V}[\hat{\theta}(I_0)]$ is an estimate of the variance of $\hat{\theta}(I_0)$. If the statistic follows asymptotically a standard normal distribution, then a confidence set with level $1-\alpha$ for the index I can be built by noting that

$$t(I_0)^2 \leq z_{\alpha/2}^2 \quad \Leftrightarrow \quad \frac{\hat{\theta}(I_0)^2}{\hat{V}[\hat{\theta}(I_0)]} \leq z_{\alpha/2}^2 \quad \Leftrightarrow \quad \hat{\theta}(I_0)^2 - z_{\alpha/2}^2\, \hat{V}[\hat{\theta}(I_0)] \leq 0\,. \tag{9}$$

This yields the confidence set

$$\mathrm{FCS}(I; 1-\alpha) = \left\{ I_0 : \hat{\theta}(I_0)^2 - z_{\alpha/2}^2\, \hat{V}[\hat{\theta}(I_0)] \leq 0 \right\}. \tag{10}$$

Since $\hat{\theta}(I_0)$ is linear in I_0, $\hat{\theta}(I_0)^2$ and $\hat{V}[\hat{\theta}(I_0)]$ are quadratic functions of I_0:

$$\hat{\theta}(I_0)^2 = A_1 I_0^2 + B_1 I_0 + C_1\,, \quad V[\hat{\theta}(I_0)] = A_2 I_0^2 + B_2 I_0 + C_2\,, \tag{11}$$

where the coefficients (defined below) depend on the data and the Gaussian critical point. On substituting (11) into (10), we get the quadratic inequality

$$A I_0^2 + B I_0 + C \leq 0 \tag{12}$$

where

$$A = A_1 - z_{\alpha/2}^2 A_2\,, \quad B = B_1 - z_{\alpha/2}^2 B_2\,, \quad C = C_1 - z_{\alpha/2}^2 C_2\,. \tag{13}$$

The coefficients, A_1, B_1, C_1, A_2, B_2, C_2 are functions of the sample moments and their variance estimates. The FCS solve the second degree polynomial inequality in (12) for I_0. Let $\Delta = B^2 - 4AC$, then the $(1-\alpha)$-level Fieller-type confidence set is characterized as follows:

1. if $\Delta > 0$ and $A > 0$, then $\mathrm{FC}(I; 1-\alpha) = \left[\frac{-B-\sqrt{\Delta}}{2A},\ \frac{-B+\sqrt{\Delta}}{2A} \right]$,
2. if $\Delta > 0$ and $A < 0$, then $\mathrm{FC}(I; 1-\alpha) = \left]-\infty,\ \frac{-B+\sqrt{\Delta}}{2A}\right] \cup \left[\frac{-B-\sqrt{\Delta}}{2A},\ +\infty \right[$,
3. if $\Delta < 0$, then $A < 0$ and $\mathrm{FC}(I; 1-\alpha) = \mathbb{R}$.

For more details, see Bolduc et al. (2010) and the references therein.

2.2 *Fieller-Type Inference for the Theil Index*

The Theil index belongs to the family of GE indices and can be written as a function of two moments $\mu = \mathbb{E}(Y)$ and $\nu = \mathbb{E}[Y \log(Y)]$, where μ and ν can be estimated using their sample counterparts. In this paper, we will use the following expression for the Theil index:

$$I_T = \frac{\nu}{\mu} - \log(\mu). \tag{14}$$

For the Theil index (I_T), the null hypothesis defined in (4) can be written as:

$$H_0(I_{T0}) : \frac{\nu}{\mu} - \log(\mu) = I_{T0}\,. \tag{15}$$

The variance of the estimated Theil index can be derived using the Delta method and it is defined by (3) where the expressions of the derivatives in this context are:

$$\frac{\partial \psi}{\partial \mu} = -\frac{(\nu + \mu)}{\mu^2}\,, \quad \frac{\partial \psi}{\partial \nu} = \frac{1}{\mu}\,. \tag{16}$$

The Fieller-type method for the Theil index starts by considering the equivalent linear hypothesis as shown in (7):

$$H_0(I_{T0}) : \nu - \mu\, \log(\mu) - \mu\, I_{T0} = 0\,, \tag{17}$$

along with the corresponding t-statistics (squared). The confidence set for I_T is then obtained by solving the quadratic inequality described by (10), (11) and (12). For this, we derive the parameters $A1$, $B1$, $C1$, $A2$, $B2$ and $C2$ in Eq. (11) for the Theil index:

$$A_1 = \hat{\mu}^2\,, \quad B_1 = -2\hat{\mu}\,[\hat{\nu} - \hat{\mu}\log(\hat{\mu})]\,, \quad C_1 = [\hat{\nu} - \hat{\mu}\log(\hat{\mu})]^2\,. \tag{18}$$

To get the variance of $\hat{\theta}(I_0)$, we apply the Delta method to $\theta(I_0)$ in (7):

$$A_2 = \hat{\sigma}^2_\mu/n\,, \quad B_2 = \big(2\hat{\sigma}^2_\mu\,[\log(\hat{\mu}) + 1] - 2\hat{\sigma}_{\mu\nu}\big)/n\,, \tag{19}$$

$$C_2 = \big(\hat{\sigma}^2_\mu[\log(\hat{\mu}) + 1]^2 - 2\hat{\sigma}_{\mu\nu}[\log(\hat{\mu}) + 1] + \hat{\sigma}^2_\mu\big)/n\,, \tag{20}$$

where $\hat{\sigma}^2_\mu$, $\hat{\sigma}^2_\nu$ and $\hat{\sigma}_{\mu\nu}$ are defined in (3).

3 Simulation Results

In this section, we provide Monte Carlo evidence on the finite-sample properties of the Fieller-type method for the Theil index. We conduct several simulation studies focusing on the behaviour of the Fieller method when the hypothesized income distribution under the null is characterized by thick tails. To this end, we simulate data sets from the Singh-Maddala distribution [$Y_i \sim \text{SM}(a, b, q)$], the Gamma distribution [Gamma(k, θ)], and finite mixtures of the latter. These distributions have been used in the literature in the context of income inequality measures (Brachmann et al. 1996; McDonald 1984; Kleiber and Kotz 2003; Cowell and Victoria-Feser 1996).

The CDF of the Singh-Maddala distribution can be written as

$$F(y) = 1 - \left[1 + \left(\frac{y}{b}\right)^a\right]^{-q}, \tag{21}$$

where a is a shape parameter which affects both tails, q is another shape parameter which affects only the right tail, and b is a scale parameter which has no impact on our analysis (for the indices in question are scale invariant. For this distribution, the k-th moment exists for $-a < k < aq$.

The expectation of the Singh-Maddala distribution can be expressed as

$$\mu = \frac{q\, b\, \Gamma\left(a^{-1}+1\right)\Gamma\left(q-a^{-1}\right)}{\Gamma\left(q+1\right)}\,. \tag{22}$$

In this case, a closed-form expression for $\nu = \mathbb{E}[Y\,\log(Y)]$ is also available:

$$\nu = \mu\, a^{-1}[\psi(a^{-1}+1) - \psi(q-a^{-1}) + a\log(b)] \tag{23}$$

where $\Gamma(\cdot)$ is the Gamma function and $\psi(\cdot) \equiv \Gamma'(\cdot)/\Gamma(\cdot)$ is the digamma function.

The other distribution we consider in the simulations is the Gamma distribution with density function

$$f(y) = \frac{y^{k-1}e^{-(k/\theta)}}{\theta^k \Gamma(k)}\,,\ y > 0\,, \tag{24}$$

where k is a shape parameter and θ is a scale parameter. The expectation of this distribution (μ) is the scale multiplied by the shape parameter ($\mu = k\theta$). The value of ν for the Gamma distribution was computed by numerical methods.

The number of replications was set to $N = 10{,}000$. For each sample, we compute the Theil inequality, and the underlying estimated variance, and the t-type statistics associated with the Delta and Fieller-type methods. Because of the duality between tests and confidence sets, the coverage rate of the confidence sets can be evaluated by computing the rejection probabilities of these tests. The coverage error rate (or equivalently the rejection probability) is computed as the proportion of times the relevant t-statistic rejects the null hypothesis. For a significance level α, we say the test approaches the nominal level when the rejection rate approaches α.

The main results of the simulation experiments are presented in the form of plots where the numbers of observations are on the x-axis and the coverage error rates on the y-axis. The 5% nominal level is maintained for all tests. The horizontal solid lines in the graphs represent the nominal level 0.05.

Our simulation results show that the Fieller-type method has better coverage than the Delta method in several cases, especially when the underlying distribution involves heavy lower and upper tails.

Figure 1 plots the rejection probabilities of the Fieller-type and Delta methods under Singh-Maddala distributions. In the left panel, the distribution is Singh-Maddala with parameters $a = 2.8$ and $q = 1.7$. The Fieller-type and Delta methods have similar coverage. However, the other designs considered reveal important improvement with the Fieller method. In the right panel, the distribution is Singh-Maddala with parameters $a = 1.1$ and $q = 5$. For this choice of parameters, the distribution exhibits bunching of low observations. In this context, the Fieller-type method outperforms the Delta method for relatively small samples up to 400 observations.

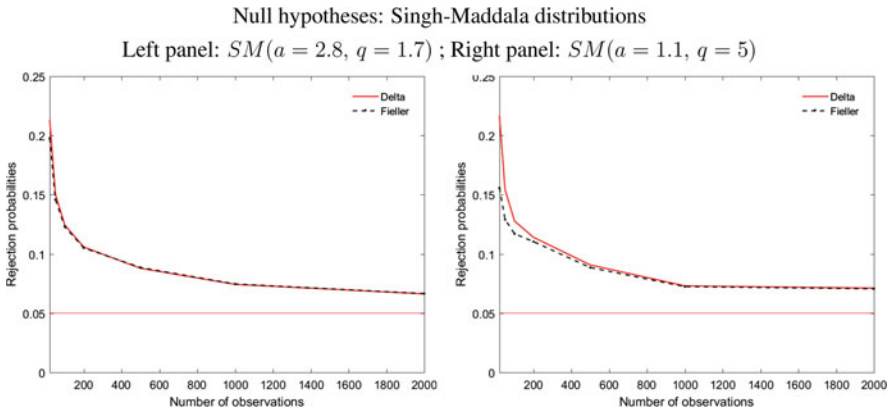


Fig. 1 Rejection probabilities for Delta and Fieller methods (Note – The Delta method and Fieller method statistics are defined by (5) and (8) respectively. The null hypothesis tested is $H(I_{T0}) : I = I_0$ where I_0 is computed analytically)

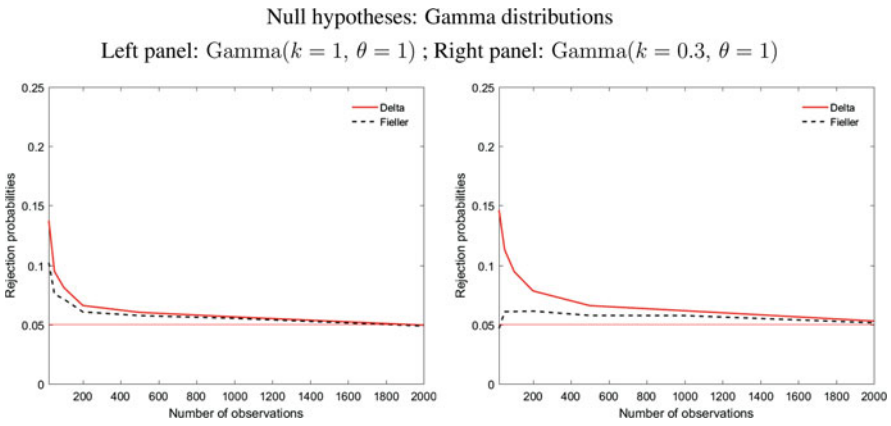


Fig. 2 Rejection probabilities for Delta and Fieller methods (Note – The Delta method and Fieller method statistics are defined by (5) and (8) respectively. The null hypothesis tested is $H(I_{T0}) : I = I_0$ where I_0 is computed analytically)

The same conclusion can be drawn from the case where we assume a Gamma distribution under the null. The size improvements we find with the Fieller-type method increase as the left tail of the distribution gets thicker. In the left panel of Fig. 2, we plot the rejection probabilities under both methods under a Gamma($k = 1$, $\theta = 1$) distribution. The differences in the rejection probabilities for samples of size 20 is around 4%, and around 2% with 100 observations. As we increase the proportion of low observations, the Fieller-type method provides remarkable size improvements, and in some cases it approaches the 5% nominal level for sample sizes as small as 200. In the right panel of this figure, the Fieller-type method coverage error is less than that of Delta method by around 9% going down from almost 15–6%.

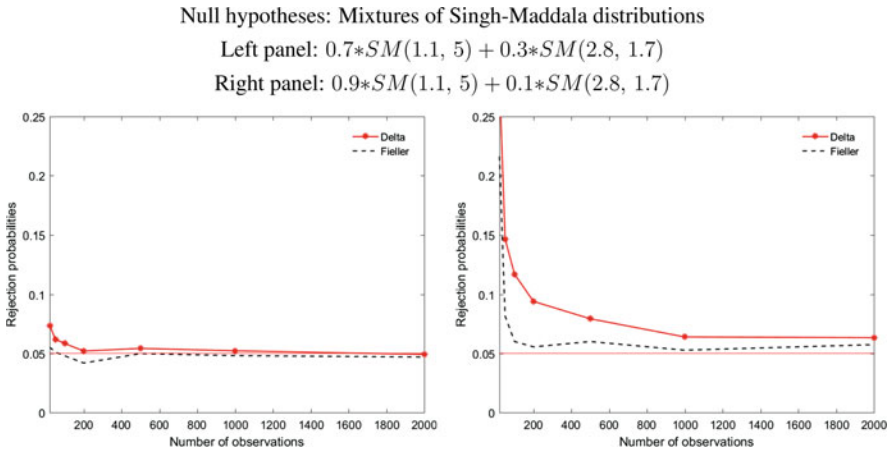


Fig. 3 Rejection probabilities for Delta and Fieller methods (Note – The Delta method and Fieller method statistics are defined by (5) and (8) respectively. The null hypothesis tested is $H(I_{T0}) : I = I_0$ where I_0 is calibrated via a separate simulation)

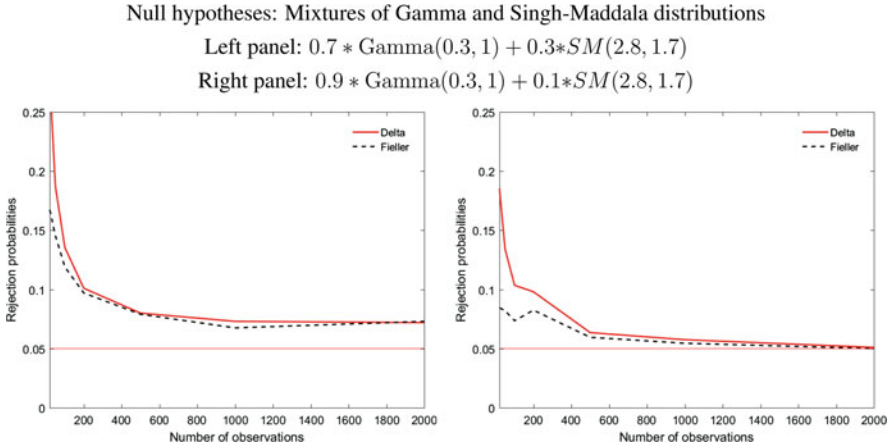


Fig. 4 Rejection probabilities for Delta and Fieller methods (Note – The Delta method and Fieller method statistics are defined by (5) and (8) respectively. The null hypothesis tested is $H(I_{T0}) : I = I_0$ where I_0 is calibrated via a separate simulation)

The shape of the distributions underlying the aforementioned results represent populations where most of the individuals are poor and few are rich. The choice of these distributions was made to study the performance of the two methods when tails are fat both near the zero boundary and to the right. As we will discuss shortly our findings conform with the theoretical work of Cowell and Victoria-Feser (1996).

In Figs. 3 and 4, we consider mixed designs with bimodal distributions under the null. These distributions, their parameters and the associated mixture weights are chosen to capture represent tail thickness at both ends of the distribution.

Since the analytical expression for the Theil under the null of mixtures does not exist, we used an estimate of the Theil index based on a very large sample (n=1000,000,000). This approach of computing the true Theil index under the null is justified by the consistency of the Theil Index.

Figure 3 plots the rejection probabilities for mixtures of two Singh-Maddala distributions. In the left panel, the mixture combines a SM(1.1, 5) with probability 0.7 weight and SM(2.8, 1.7) with probability 0.3: i.e. on average, we draw 70% of the sample from a distribution with a peak near low incomes, and 30% from a distribution characterized by a thick right tail. For this design, the Fieller-type method approaches the nominal significance level for samples as small as 50 observations.

In the right panel, we increase the weight for the first distribution from 0.7 to 0.9. Thus we are giving more weight to the distribution with irregularities on the left tail rather than the one characterized with right tail thickness. The Fieller-type method dominates the Delta method by wide margins. This confirms our previous conclusion that the Fieller-type method is superior to the Delta method, especially when the distribution exhibits bunching of low observations.

Further evidence appears in Fig. 4 where we consider mixtures of Gamma(0.3, 1) and SM(2.8, 1.7) distributions. Again the left panel gives to the first distribution (the Gamma distribution) a weight of 0.7, while the right one increases this weight to 0.9. Again, the Fieller-type method improves coverage, especially for small samples.

The designs we considered can be interpreted through the work of Cowell and Victoria-Feser (1996). This paper views the underlying distribution as a mixture of a finite number of other distributions where bunching and tail behaviour can be formally modelled. Results for cases of "extreme" behaviour are pointed out, including the zero and infinite boundaries. In particular, the Theil index can have an unbounded influence function when some of the data approaches ∞, although the zero boundary may matter for other inequality indices.

The influence function measures the change in the estimator for a small perturbation of the data. It is related to the bias of the estimator in the sense that when the IF is unbounded the bias can be infinite. Cowell and Victoria-Feser (1996) show that any decomposable scale invariant index for which the mean is estimated from the sample has an unbounded *IF*. We find that the Fieller-type method improves coverage, especially when small (near zero) or large observations are highly probable.

4 Conclusion

This paper proposes Fieller-type procedures for inference on the Theil inequality index and illustrates its superiority relative to its standard Delta method counterpart, using various empirically relevant simulation designs. Our results confirm that, in contrast with the Delta method, the proposed procedures can capture some of the distributional irregularities arising from the concentration of low observations and

the thickness of the right tail. More broadly, our findings suggest that the Fieller-type approach holds concrete promise for many other inequality measures, as well as for inference on differences between measures.

References

Bennett, B. (1953). Some further extensions of Fieller's theorem. *Annals of the Institute of Statistical Mathematics, 5*(1), 103–106.

Bennett, B. (1959). On a multivariate version of Fieller's theorem. *Journal of the Royal Statistical Society Series B, 21*(1), 59–62.

Bernard, J. T., Idoudi, N., Khalaf, L., & Yélou, C. (2007). Finite sample inference methods for dynamic energy demand models. *Journal of Applied Econometrics, 22*(7), 1211–1226.

Bernard, B., Chu, B., Khalaf, L., & Voia, M. (2015). *Non-standard confidence sets for ratios and tipping points with applications to dynamic panel data*. Paper presented at the International Panel Data Conference, Budapest.

Biewen, M. (2002). Bootstrap inference for inequality, mobility and poverty measurement. *Journal of Econometrics, 108*, 317–342.

Bolduc, D., Khalaf, L., & Yélou, C. (2010). Identification robust confidence set methods for inference on parameter ratios with application to discrete choice models. *Journal of Econometrics, 157*, 317–327.

Brachmann, K., Stich, A., & Trede, M. (1996). Evaluating parametric income distribution models. *Allgemeines Statistisches Archiv, 80*, 285–298.

Cowell, F., & Flachaire, E. (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics, 141*, 1044–1072.

Cowell, F., & Flachaire, E. (2015). Statistical methods for distributional analysis. In A. B. Atkinson & F. Bourguignon (Eds.), *Handbook of income distribution* (Vol. 2, pp.359–465). Oxford: North Holland.

Cowell, F. A., & Victoria-Feser, M. P. (1996). Robustness properties of inequality measures. *Econometrica, 64*(1), 77–101.

Cox, D. (1967). Fieller's theorem and a generalization. *Biometrika, 54*, 567–572.

Davidson, R., & Flachaire, E. (2007). Asymptotic and bootstrap inference for inequality and poverty measures. *Journal of Econometrics, 141*, 141–166.

Dufour, J. -M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica, 65*, 1365–1387.

Dufour, J. -M, Flachaire, E., & Khalaf, L. (2017). Permutation tests for comparing inequality measures. *Journal of Business & Economic Statistics*, (just-accepted).

Fieller E. C. (1940). The biological standardization of insulin. *Journal of the Royal Statistical Society (Supplement), 7*, 1–64.

Fieller E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society B, 16*, 175–185.

Gleser, L. J., & Hwang, J. T. (1987). The nonexistence of $100(1 - \alpha)$ confidence sets of finite expected diameter in errors-in-variables and related models. *The Annals of Statistics, 15*, 1351–1362.

Johannesson, M., Jönsson, B., & Karlsson, G. (1996). Outcome measurement in economic evaluation. *Health Economics, 5*, 279–296.

Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. Hoboken: Wiley.

Laska, E., Meisner, M., & Siegel, C. (1997). Statistical inference for cost effectiveness ratios. *Health Economics, 6*, 229–242.

Maasoumi, E. (1997). Empirical analyses of inequality and welfare. In M. H. Pesaran, M. Wickens, & P. Schmidt (Eds.), *Handbook of applied econometrics: Microeconomics* (Vol. 2, pp. 202–245). Oxford: Blackwell

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica, 52*, 647–664.

Mills, J., & Zandvakili, S. (1997). Statistical inference via bootstrapping for measures of inequality. *Journal of Applied Econometrics, 12*, 133–150.

Scheffé, H. (1970). Multiple testing versus multiple estimation. Improper confidence sets. Estimation of directions and ratios. *The Annals of Mathematical Statistics, 41*(1), 1–29.

Srivastava, M. S. (1986). Multivariate bioassay, combination of bioassays, and Fieller's theorem. *Biometrics, 42*, 131–141.

Stock, J., & Lazarus, E. (2016). *Identification of factor-augmenting technical growth and the decline of the labor share*. Paper presented at the CIREQ Econometrics Conference in Honor of Jean-Marie Dufour, Montréal.

von Luxburg, U., & Franz, V. (2004). Confidence sets for ratios: A purely geometric approach to Fieller's theorem. Technical report TR-133, Max Planck Institute for Biological Cybernetics.

Willan, A., & O'Brien, B. (1996). Confidence intervals for cost effectiveness ratios: An application of Fieller's theorem. *Health Economics, 5*, 297–305.

Zerbe, G. (1978). On Fieller's theorem and the general linear model. *The American Statistician, 32*, 103–105.

Zerbe, G. O., Laska, E., Meisner, M., & Kushner, H. B. (1982). On multivariate confidence regions and simultaneous confidence limits for ratios. *Communications in Statistics – Theory and Methods, 11*, 2401–2425.

Poverty-Dominant Marginal Transfer Reforms in Socially Risky Situations

Paul Makdissi and Quentin Wodon

Abstract Public transfer programs have a dual objective of redistributing income and providing insurance when individual incomes are subject to variability. To our knowledge, the normative literature on marginal policy reform has not yet provided a method to account for individual's exposure to risk in the evaluation of public transfer programs. Stochastic dominance tests are proposed to identify robust marginal poverty-reducing transfer reforms in socially risky situations.

Keywords Poverty · Policy reform · Risk · Stochastic dominance

JEL Codes: H23, I32

1 Introduction

When constructing poverty comparisons or analyzing the impact of marginal policy reforms on poverty, economists usually rely on cross-sectional data which give information on income at one point in time. However, it has long been recognized that incomes are subject to uncertainty and that this may affect welfare if individuals are risk averse and cannot insure properly. Indeed, many transfer programs are especially designed to enable households to better cope with negative income shocks. This is the case in OECD countries for unemployment benefits, among others. In developing countries, safety nets often take the form of public works which are expanded during recessions and phased out, or at least reduced during booms.

P. Makdissi (✉)
Department of Economics, University of Ottawa, Ottawa, ON, Canada
e-mail: paul.makdissi@uottawa.ca

Q. Wodon
Education Global Practice, World Bank, Washington, DC, USA
e-mail: qwodon@worldbank.org

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_7

While such transfers are designed to cope with risk, the analysis of their impact on poverty and social welfare is typically based on cross-sectional household survey data that provide income or consumption information for one period of time only. With the progressively wider availability of panel data on households, even in developing countries, it becomes feasible to directly take into account uncertainty in poverty comparisons and policy or program reform evaluations.

In this paper, we use the standard concept of a household's certainty equivalent income (which depends on both expected income and income variability) in order to develop stochastic dominance tests for poverty comparisons and for evaluation of the impact on poverty of marginal policy reforms in socially risky situations. In doing so, we build on two strands of the literature. We first follow Gravel and Tarroux (2015) and show how to use in a risk framework the results on robust poverty ordering tests developed among others by Atkinson (1987), Foster and Shorrocks (1988a,b), and Zheng (1999), and extended to multidimensional comparisons by Duclos et al. (2006) and Bourguignon and Chakravarty (2002). Our proposition differs from Gravel and Tarroux (2015) by allowing for a continuum of risk levels. Next, building on Makdissi and Wodon (2002) and Duclos et al. (2005a,b, 2008) who extended the work on the impact of marginal policy reforms by Yitzhaki and Slemrod (1991) and Mayshar and Yitzhaki (1995), we show how to integrate uncertainty in the evaluation of marginal policy reforms in socially risky situations.

Section 2 of ther paper provides our framework for robust poverty comparisons. Section 3 deals with robust policy reform orderings. A brief conclusion follows.

2 Robust Poverty Comparisons

In this section, we adapt Duclos et al. (2006) multidimensional stochastic dominance test to the context of income and risk. We also extend their result to higher order of dominance.

Income poverty is measured using a bivariate distribution of expected income, μ, and income variance, σ^2, drawn from the set $\Im$:

$$\Im := \{F : [0, a] \times [0, v] \to [0, 1] \,|F \text{ is nondecreasing, continuous and onto}\}, \quad (1)$$

where F is the bivariate cumulative distribution of expected income and income variance, and a and v are values equal to or exceeding the maximum conceivable expected income and variance, respectively. Theoretically, those values are the support of the data generating process underlying F. In an empirical application the values can be chosen to excess substantially all observations in the data set. As long as the chosen value are finite, the exact values do not play a role in the dominance tests derived below. Under risk aversion, poverty measures should be based on a household's certainty equivalent income rather than on expected income. For simplicity, we assume that income variability represents risk and that

households cannot insure. The poverty line function $z\left(\sigma^2\right)$ identifies a household as poor if its expected income falls under $z\left(\sigma^2\right)$. The poverty index $P\left(z\left(\cdot\right)\right)$ used to aggregate household level contributions to poverty is additive such that $P\left(z\left(\cdot\right)\right) = \int_0^v \int_0^a p\left(\mu,\sigma^2\right) f\left(\mu,\sigma^2\right) d\mu d\sigma^2$ where f is the density function associated with F. This implies that $F\left(\mu,\sigma^2\right) = \int_0^{\sigma^2} \int_0^{\mu} f\left(x,y\right) dxdy$, $p\left(\mu,\sigma^2\right) \geq 0$ for all μ and σ^2 and $p\left(\mu,\sigma^2\right) = 0$ if $\mu \geq \sigma^2$. The function $p\left(\mu,\sigma^2\right)$ is the contribution to total poverty of a household with expected income μ and variance σ^2. For expositional simplicity and ease of proofs, we define income security, ξ, as $\xi = v - \sigma^2$, and rewrite the poverty line function as $\phi\left(\xi\right) = z\left(v - \xi\right)$. We assume that $\phi'\left(\xi\right) \leq 0$ so that if uncertainty decreases or income security increases, the household needs a lower expected income in order not to be poor. Let $\widetilde{F} \in \Im$ be the bivariate cumulative distribution of (μ,ξ) such that $\widetilde{F}(\mu,\xi) = F(\mu, v-\xi)$ and let $\widetilde{f}$ be the density function associated with $\widetilde{F}$. The poverty indices are

$$\widetilde{P}\left(\phi \quad \left(\cdot\right)\right) = \int_0^v \int_0^a \widetilde{p}\left(\mu,\xi\right)\widetilde{f}\left(\mu,\xi\right) d\mu d\xi, \tag{2}$$

with

$$\left.\begin{array}{c} \widetilde{p}\left(\mu,\xi\right) \geq 0 \text{ forall} \mu \text{and} \xi \\ \widetilde{p}\left(\mu,\xi\right) = 0 \text{ if} \mu \geq \phi\left(\xi\right) \end{array}\right\}. \tag{3}$$

These indices are classified into classes Π^s such that

$$\Pi^s\left(\phi\left(\cdot\right)\right) = \left\{ \widetilde{P}\left(\cdot\right) \left| \begin{array}{c} \widetilde{p}\left(\mu,\xi\right) \geq 0 \forall \mu, \forall \xi, \widetilde{p}\left(\mu,\xi\right) \in C^s \\ \widetilde{p}\left(\mu,\xi\right) = 0 \text{if} \mu \geq \phi\left(\xi\right) \\ (-1)^i \widetilde{p}^{\,(i,0)} \geq 0, \widetilde{p}^{(0,1)} \leq 0, (-1)^i \widetilde{p}^{(i,1)} \leq 0 \end{array} \right. \right\}, \tag{4}$$

where C^s is the set of continuous and piecewise differentiable functions over $[0,a] \times [o,v]$, and $\widetilde{p}^{(i,j)}$ represents the $(i+j)$th partial derivative with i times with respect to μ and j times with respect to ξ. The Foster et al. (1984) poverty index corresponding to any given value of α

$$FGT^{\alpha}\left(\phi\left(\cdot\right)\right) = \int_0^v \int_0^{\phi(\xi)} \left[\left(\phi\left(\xi\right) - \mu\right)/\phi\left(\xi\right)\right]^{\alpha} \widetilde{f}\left(\mu,\xi\right) d\mu d\xi \tag{5}$$

belongs to the class $s = \alpha + 1$.

The definition of $\Pi^1\left(\phi\left(\cdot\right)\right)$ is such that an increase of expected income cannot increase poverty for a given level of income security. It also implies that an increase in income security cannot increase poverty whatever the expected income. Note that the poverty reduction through a one dollar increase in expected income is larger at lower levels of income security, an assumption similar to Sen (1997) weak equity axiom.

For $\Pi^2\left(\phi\left(\cdot\right)\right)$, the definition (4) says that an equalizing transfer of \$1 to a poor from a richer individual with the same income security decreases poverty, and this effect is stronger across individuals at lower levels of income security. The normative interpretation of (4) for higher s can be made using Fishburn

and Willig (1984), whose general transfer principles give increasing weights to transfers occurring at the bottom of the distribution as s increases. Again, (4) makes the associated poverty reduction larger for households at lower levels of income security.[1]

Let $\Delta\Theta_{AB} = \Theta_B - \Theta_A$ for any function Θ, and define:

$$D^s(\mu,\xi) = \begin{cases} \widetilde{F}(\mu,\xi) & \text{for} s = 1 \\ \int_0^\mu D^{s-1}(\mu,\xi)\, d\mu & \text{for} s \in \{2,3,4,\ldots\} \end{cases}. \tag{6}$$

Proposition 1 $\Delta\widetilde{P}_{AB} \leq 0$ *for all poverty indices* $\widetilde{P} \in \Pi^s(\phi(\cdot))$ *and for all poverty line functions* $\phi(\xi) \leq \phi^+(\xi)\ \forall\xi$ *if*

$$\Delta D^s_{AB}(\mu,\xi) \leq 0 \forall \mu \in \left[0, \phi^+(\xi)\right] \wedge \forall \xi \in [0, v]. \tag{7}$$

Proof Integrating by parts equation (2) with respect to μ gives

$$\begin{aligned} \widetilde{P}(\phi(\cdot)) = &\int_0^v \widetilde{p}(\mu,\xi) F(\mu|\xi)\big|_0^{\phi(\xi)} \widetilde{f}(\xi)\, d\xi \\ &- \int_0^v \int_0^{\phi(\xi)} \widetilde{p}^{(1,0)}(\mu,\xi) \widetilde{F}(\mu|\xi) \widetilde{f}(\xi)\, d\mu d\xi. \end{aligned} \tag{8}$$

Since $\widetilde{F}(\mu = 0|\xi) = 0$ and $\widetilde{p}(\phi(\xi),\xi) = 0$, the first term on the right hand side of the equation equals 0. Define $I(\xi) = \int_0^{\phi(\xi)} \widetilde{p}^{(1,0)}(\mu,\xi) \widetilde{F}(\mu,\xi)\, d\mu$ and note that

$$\begin{aligned} \frac{dI(\xi)}{d\xi} = &\phi'(\xi) \widetilde{p}^{(1,0)}(\phi(\xi),\xi) \widetilde{F}(\phi(\xi),\xi) \\ &+ \int_0^{\phi(\xi)} \widetilde{p}^{(1,1)}(\mu,\xi) \widetilde{F}(\mu,\xi)\, d\mu \\ &+ \int_0^{\phi(\xi)} \widetilde{p}^{(1,0)}(\mu,\xi) \widetilde{F}(\mu|\xi) \widetilde{f}(\xi)\, d\mu. \end{aligned} \tag{9}$$

Integrating (9) from 0 to v gives

$$\begin{aligned} \int_0^v \int_0^{\phi(\xi)} \widetilde{p}^{(1,0)}(\mu,\xi) \widetilde{F}(\mu|\xi) \widetilde{f}(\xi)\, d\mu d\xi = &\int_0^v \phi'(\xi) \widetilde{p}^{(1,0)}(\phi(\xi),\xi) \widetilde{F}(\phi(\xi),\xi)\, d\xi \\ &+ \int_0^v \int_0^{\phi(\xi)} \widetilde{p}^{(1,1)}(\mu,\xi) \widetilde{F}(\mu,\xi)\, d\mu d\xi \\ &+ \int_0^v \int_0^{\phi(\xi)} \widetilde{p}^{(1,0)}(\mu,\xi) \widetilde{F}(\mu|\xi) \widetilde{f}(\xi)\, d\mu d\xi. \end{aligned} \tag{10}$$

[1] For each order s, we have the standard Fishburn and Willig normative interpretation of s-order unidimensional dominance (that is, the interpretation of $(-1)^i \widetilde{p}^{(i,0)} \geq 0$), joined with a weak version of the traditional normative interpretation of $s+1$-order dominance (the interpretation of $(-1)^i \widetilde{p}^{(i,1)} \leq 0$).

Using Eq. (10), we can rewrite Eq. (8) as

$$\begin{aligned}\widetilde{P}(\phi(\cdot)) = &-\int_0^{\phi(v)} \widetilde{p}^{(1,0)}(\mu,\xi)\,\widetilde{F}(\mu,\xi)\,d\mu \\ &+\int_0^{v} \phi'(\xi)\,\widetilde{p}^{(1,0)}(\phi(\xi),\xi)\,\widetilde{F}(\phi(\xi),\xi)\,d\xi \\ &+\int_0^{v}\int_0^{\phi(\xi)} \widetilde{p}^{(1,1)}(\mu,\xi)\,\widetilde{F}(\mu,\xi)\,d\mu d\xi.\end{aligned} \tag{11}$$

The definition of $\Pi^1(\phi(\cdot))$ implies that the second term on the right hand side of the equation is equal to 0. Noting that $D^1(\mu,\xi) = \widetilde{F}(\mu,\xi)$ we can rewrite (11) as

$$\begin{aligned}\widetilde{P}(\phi(\cdot)) = &-\int_0^{\phi(v)} \widetilde{p}^{(1,0)}(\mu,\xi)\,D^1(\mu,\xi)\,d\mu \\ &+\int_0^{v}\int_0^{\phi(\xi)} \widetilde{p}^{(1,1)}(\mu,\xi)\,D^1(\mu,\xi)\,d\mu d\xi.\end{aligned} \tag{12}$$

Now assume that for $s-1$ we have

$$\begin{aligned}\widetilde{P}(\phi(\cdot)) = &(-1)^{s-1}\int_0^{\phi(v)} \widetilde{p}^{(s-1,0)}(\mu,\xi)\,D^{s-1}(\mu,\xi)\,d\mu \\ &+(-1)^{s}\int_0^{v}\int_0^{\phi(\xi)} \widetilde{p}^{(s-1,1)}(\mu,\xi)\,D^{s-1}(\mu,\xi)\,d\mu d\xi\end{aligned} \tag{13}$$

Integrating (13) by parts with respect to μ gives

$$\begin{aligned}\widetilde{P}(\phi(\cdot)) = &(-1)^{s-1}\,\widetilde{p}^{(s-1,0)}(\mu,\xi)\,D^{s}(\mu,\xi)\Big|_0^{\phi(v)} \\ &-(-1)^{s-1}\int_0^{\phi(v)} \widetilde{p}^{(s,0)}(\mu,\xi)\,D^{s}(\mu,\xi)\,d\mu \\ &+(-1)^{s}\int_0^{v} \widetilde{p}^{(s-1,1)}(\mu,\xi)\,D^{s}(\mu,\xi)\,d\xi\Big|_0^{\phi(v)} \\ &-(-1)^{s}\int_0^{v}\int_0^{\phi(\xi)} \widetilde{p}^{(s,1)}(\mu,\xi)\,D^{s}(\mu,\xi)\,d\mu d\xi.\end{aligned} \tag{14}$$

The first and third terms on the right hand side of the equation are equal to 0. Equation (14) than be rewritten as

$$\begin{aligned}\widetilde{P}(\phi(\cdot)) = &(-1)^{s}\int_0^{\phi(v)} \widetilde{p}^{(s,0)}(\mu,\xi)\,D^{s}(\mu,\xi)\,d\mu \\ &+(-1)^{s+1}\int_0^{v}\int_0^{\phi(\xi)} \widetilde{p}^{(s,1)}(\mu,\xi)\,D^{s}(\mu,\xi)\,d\mu d\xi.\end{aligned} \tag{15}$$

Equations (12) and (15) have the same structure than Eq. (13). This implies that Eq. (15) is true for every $s \in \{1, 2, 3, \ldots\}$ Using (15), we can write

$$\Delta\widetilde{P}_{AB} = (-1)^s \int_0^{\phi(v)} \widetilde{p}^{(s,0)}(\mu, \xi)\, \Delta D^s_{AB}(\mu, \xi)\, d\mu + (-1)^{s+1} \int_0^{v} \int_0^{\phi(\xi)} \widetilde{p}^{(s,1)}(\mu, \xi)\, \Delta D^s_{AB}(\mu, \xi)\, d\mu d\xi \tag{16}$$

The definition of $\Pi^s(\phi(\cdot))$ and Eq. (16) imply that Proposition 1 is true. ■

Essentially, the graphical analysis of $\Delta D^s_{AB}(\mu, \xi)$ implies the comparison of surfaces to make sure that they do not intersect. This enables us to make robust bi-dimentional poverty comparisons in a way similar to Duclos et al. (2006) except that we have extended their result for all orders of stochastic dominance.

3 Robust Policy Reform Orderings

In addition to their redistributive objective, transfer policies often have a dual role of providing insurance for income shocks. Any change in these policies affects both individuals' expected incomes and income variability. In this section, we use the dominance surface of the observed distribution as an anchor point and analyze how a marginal reform of transfer policies affects it. This allows us to derive a dominance condition for policy reforms focussing on both channels of transmission: expected income and income variability.

We now turn to the analysis of the impact on poverty of marginal policy reforms. Household income is the sum of k income sources, some of them being public transfers, so that expected total income is $\mu = \sum_{i=1}^{k} \mu_i$, where μ_i is the expected income from source i at total expected income level μ. Using the definition of the variance of income $\sigma^2 = \sum_{i=1}^{k} \sigma_i^2 + 2\sum_{i=1}^{k} \sum_{j=i+1}^{k} COV_{i,j}$, where $COV_{i,j}$ is the covariance between income source i and income source j, we can write

$$\xi = v - \sum_{i=1}^{k} \sigma_i^2 - \sum_{i=1}^{k} COV_{i,-i}, \tag{17}$$

where $COV_{i,-i}$ represents the covariance between income source i and all other sources of income. The impact of a proportional marginal increase of income from source i on $\widetilde{p}(\mu, \xi)$ is

$$\begin{aligned} d\widetilde{p}(\mu, \xi) &= \left[\widetilde{p}^{(1,0)}(\mu, \xi)\, \frac{\partial \mu}{\partial t_i} + \widetilde{p}^{(0,1)}(\mu, \xi)\, \frac{\partial \xi}{\partial t_i}\right] dt_i \\ &= \left[\widetilde{p}^{(1,0)}(\mu, \xi)\, t_i(\mu, \xi) + \widetilde{p}^{(0,1)}(\mu, \xi) \left(-\sigma_i^2(\mu, \xi) - COV_{i,-i}(\mu, \xi)\right)\right] dt_i, \end{aligned} \tag{18}$$

where $t_i(\mu,\xi)$, $\sigma_i^2(\mu,\xi)$ and $COV_{i,-i}(\mu,\xi)$ represent respectively the expected income or transfer from source i, the variance of transfers from that source and the covariance between transfers from that source and other sources of incomes for a household with expected income μ and income security ξ.

The impact on $\widetilde{p}(\mu,\xi)$ of a marginal policy reform that reduces transfers from income source k and increases transfers from income source l is

$$d\widetilde{p}(\mu,\xi) = \left[\widetilde{p}^{(1,0)}(\mu,\xi)\, t_l(\mu,\xi) + \widetilde{p}^{(0,1)}(\mu,\xi)\left(-\sigma_l^2(\mu,\xi) - COV_{l,-l}(\mu,\xi)\right)\right] dt_l \quad (19)$$
$$+ \left[\widetilde{p}^{(1,0)}(\mu,\xi)\, t_k(\mu,\xi) + \widetilde{p}^{(0,1)}(\mu,\xi)\left(-\sigma_k^2(\mu,\xi) - COV_{k,-k}(\mu,\xi)\right)\right] dt_k.$$

The impact of the reform on the government budget B is

$$dB = \frac{\partial B}{\partial t_l} dt_l + \frac{\partial B}{\partial t_k} dt_k. \quad (20)$$

Under budget neutrality, $dB = 0$. Following Duclos et al. (2005a), we define the economic efficiency ratio for the reform, γ, as

$$\gamma = \frac{(\partial B/\partial t_l)/T_l}{(\partial B/\partial t_k)/T_k}, \quad (21)$$

where $T_l = \int_0^v \int_0^a t_l(\mu,\xi)\widetilde{f}(\mu,\xi)\, d\mu d\xi$ represents the average expected transfer from source l in the population and T_k is defined analogously. The numerator and denominator in (21) give the budgetary costs per dollar of increasing household through sources k and l. Thus, γ takes into account potential differences in the marginal cost of fund of the two transfers.

Using Eqs. (21) and (20), we can rewrite (19) as

$$d\widetilde{p}(\mu,\xi) = \left\{\widetilde{p}^{(1,0)}(\mu,\xi)\left[\frac{t_l(\mu,\xi)}{T_l} - \gamma\frac{t_k(\mu,\xi)}{T_k}\right]T_l \right. \quad (22)$$
$$-\widetilde{p}^{(0,1)}(\mu,\xi)\left[\frac{\sigma_l^2(\mu,\xi)}{T_l} - \gamma\frac{\sigma_k^2(\mu,\xi)}{T_k}\right]T_l$$
$$\left. -\widetilde{p}^{(0,1)}(\mu,\xi)\left[\frac{COV_{l,-l}(\mu,\xi)}{T_l} - \gamma\frac{COV_{k,-k}(\mu,\xi)}{T_k}\right]T_l\right\} dt_l.$$

The impact of the reform on poverty is

$$d\widetilde{P}(\phi(\cdot)) = \left\{T_l\int_0^v\int_0^a \widetilde{p}^{(1,0)}(\mu,\xi)\left[\frac{t_l(\mu,\xi)}{T_l} - \gamma\frac{t_k(\mu,\xi)}{T_k}\right]\widetilde{f}(\mu,\xi)\, d\mu d\xi \right. \quad (23)$$
$$-2T_l\int_0^v\int_0^a \widetilde{p}^{(0,1)}(\mu,\xi)\left[\frac{\sigma_l^2(\mu,\xi)}{T_l} - \gamma\frac{\sigma_k^2(\mu,\xi)}{T_k}\right]\widetilde{f}(\mu,\xi)\, d\mu d\xi$$
$$\left. -T_l\int_0^v\int_0^a \widetilde{p}^{(0,1)}(\mu,\xi)\left[\frac{COV_{l,-l}(\mu,\xi)}{T_l} - \gamma\frac{COV_{k,-k}(\mu,\xi)}{T_k}\right]\widetilde{f}(\mu,\xi)\, d\mu d\xi\right\} dt_l.$$

Using (5) and (23) we can find the impact of this reform on the *FGT* indices using

$$\widetilde{p}^{(1,0)}(\mu,\xi) = -\frac{\alpha}{\phi(\xi)}\left(\frac{(\phi(\xi)-\mu)}{\phi(\xi)}\right)^{\alpha-1} \tag{24}$$

and

$$\widetilde{p}^{(0,1)}(\mu,\xi) = \alpha\left(\frac{(\phi(\xi)-\mu)}{\phi(\xi)}\right)^{\alpha-1}\frac{\phi'(\xi)\,\mu}{(\phi(\xi))^2}. \tag{25}$$

Returning to the dominance surfaces $D^s(\mu,\xi)$ for $s \in \{2,3,4,\ldots\}$, and noting that $D^s(\mu,\xi) = ((s-1)!)^{-1}\int_0^{\xi}\int_0^{\mu}(\mu-x)\widetilde{f}(x,y)\,dxdy$, the impact of a reform reducing at the margin the resources devoted to source k and increasing marginally the resources allocated to source l leads to

$$\begin{aligned} dD^s(\mu,\xi) = &\left[\frac{\partial D^s(\mu,\xi)}{\partial\mu}t_l(\mu,\xi) + \frac{\partial D^s(\mu,\xi)}{\partial\xi}\left(-\sigma_l^2(\mu,\xi) - COV_{l,-l}(\mu,\xi)\right)\right]dt_l \\ &+\left[\frac{\partial D^s(\mu,\xi)}{\partial\mu}t_k(\mu,\xi) + \frac{\partial D^s(\mu,\xi)}{\partial\xi}\left(-\sigma_k^2(\mu,\xi) - COV_{k,-k}(\mu,\xi)\right)\right]dt_k, \end{aligned} \tag{26}$$

where

$$\begin{aligned} \frac{\partial D^s(\mu,\xi)}{\partial\mu} &= \frac{1}{(s-2)!}\int_0^{\xi}\int_0^{\mu}(\mu-x)^{s-2}\widetilde{f}(x,y)\,dxdy \\ &= D^{s-1}(\mu,\xi), \end{aligned} \tag{27}$$

and

$$\begin{aligned} \frac{\partial D^s(\mu,\xi)}{\partial\xi} &= \frac{1}{(s-1)!}\int_0^{\mu}(\mu-x)^{s-1}\widetilde{f}(x,\xi)\,dx. \\ &= D^s(\mu|\xi) \end{aligned} \tag{28}$$

Using (27) and (28), we can rewrite (26) as

$$\begin{aligned} dD^s(\mu,\xi) = &\left[D^{s-1}(\mu,\xi)\,t_l(\mu,\xi) + D^s(\mu|\xi)\left(-\sigma_l^2(\mu,\xi) - COV_{l,-l}(\mu,\xi)\right)\right]dt_l \\ &+\left[D^{s-1}(\mu,\xi)\,t_k(\mu,\xi) + D^s(\mu|\xi)\left(-\sigma_k^2(\mu,\xi) - COV_{k,-k}(\mu,\xi)\right)\right]dt_k. \end{aligned} \tag{29}$$

Using Eqs. (21) and (20), we can rewrite (29) as

$$dD^s(\mu,\xi) = \left\{D^{s-1}(\mu,\xi)\left[\frac{t_l(\mu,\xi)}{T_l} - \gamma\frac{t_k(\mu,\xi)}{T_k}\right]T_l \right. \tag{30}$$

$$-D^s(\mu|\xi)\left[\frac{\sigma_l^2(\mu,\xi)}{T_l}-\gamma\frac{\sigma_k^2(\mu,\xi)}{T_k}\right]T_l$$
$$-D^s(\mu|\xi)\left[\frac{COV_{l,-l}(\mu,\xi)}{T_l}-\gamma\frac{COV_{k,-k}(\mu,\xi)}{T_k}\right]T_l\Bigg\}\,dt_l.$$

We now introduce the concept of Program Dominance Surfaces. For each order of stochastic dominance, two pairs of dominance surfaces must be compared. The first pair relates to mean dominance (*MD*) and is used to assess the impact of a reform on expected total income. The second pair relates to variance dominance (*VD*) and is used to assess the impact of a reform on income variance. The surfaces are defined as

$$MD_i^s(\mu,\xi)=\begin{cases}\frac{t_i(\mu,\xi)}{T_i}\widetilde{f}(\mu,\xi) & \text{for} s=1\\ \frac{t_i(\mu,\xi)}{T_i}D^{s-1}(\mu,\xi) & \text{for} s\in\{2,3,4,\ldots\}\end{cases}. \tag{31}$$

$$VD_i^s(\mu,\xi)=\begin{cases}\left[\frac{\sigma_i^2(\mu,\xi)}{T_i}+\frac{COV_{i,-i}(\mu,\xi)}{T_i}\right]\widetilde{f}(\mu,\xi) & \text{for} s=1\\ \left[\frac{\sigma_i^2(\mu,\xi)}{T_i}+\frac{COV_{i,-i}(\mu,\xi)}{T_i}\right]D^s(\mu|\xi) & \text{for} s\in\{2,3,4,\ldots\}\end{cases}. \tag{32}$$

Note that $MD_i^1(\mu,\xi)$ gives the density of public spending on source i spent on households with the corresponding expected income and income security, divided by the average public spending on source i. The interpretation of $VD_i^1(\mu,\xi)$ is less straightforward: it gives the density of extra income security generated by public spending on source i on households with the corresponding income and income security, again divided by the average public spending on source i. Using (23), (31), and (32), we can now state a second result.

Proposition 2 *A revenue-neutral marginal policy reform that increases proportionately all transfers under source l and reduces proportionately all those under source k will reduce poverty for all poverty indices* $\widetilde{P}\in\Pi^s(\phi(\cdot))$ *and all poverty line functions* $\phi(\xi)\le\phi^+(\xi)\ \forall\xi\in[0,v]$ *if*

$$MD_l^s(\mu,\xi)-\gamma MD_k^s(\mu,\xi)\ge 0\forall\mu\in\left[0,\phi^+(\xi)\right]\wedge\forall\xi\in[0,v] \tag{33}$$

and

$$VD_l^s(\mu,\xi)-\gamma VD_k^s(\mu,\xi)\le 0\forall\mu\in[0,\phi(\xi)]\wedge\forall\xi\in[0,v]. \tag{34}$$

Proof For $s=1$, dt_l being negative, Eq. (23) proves Proposition 2. For $s\in\{2,3,4,\ldots\}$ we refer to Proposition 1 which implies that $d\widetilde{P}\le 0$ for all poverty indices $\widetilde{P}\in\Pi^s(\phi(\cdot))$ and for all poverty line functions $\phi(\xi)\le\phi^+(\xi)\ \forall\xi$ if

$$dD^s(\mu,\xi)\le 0\forall\mu\in\left[0,\phi^+(\xi)\right]\wedge\forall\xi\in[0,v]. \tag{35}$$

Equations (30) and (35) prove Proposition 2 for $s\in\{2,3,4,\ldots\}$.■

Comparing Proposition 2 with the unidimensional program dominance result in Duclos et al. (2005a), we see that we have two conditions here instead of one, and each condition is stated in terms of the comparison of surfaces rather than curves. Condition (33) is similar to the program dominance condition in Duclos et al. (2005a) except that we must test the condition for every level of income security. In addition, we have an additional test (34). At the first order of dominance, the two conditions mean that in order to be poverty-efficient in a robust way, a policy reform must increase expected income and increase income security for poor households.

Note also that the *VD* surfaces are additive and can be decomposed into the source's own variance dominance surfaces, corresponding to the term $\frac{\sigma_i^2(\mu,\xi)}{T_i}$ in (32), and the source's covariance dominance surfaces, corresponding to the term $\frac{COV_{i,-i}(\mu,\xi)}{T_i}$. In a situation in which test (33) succeeds but test (34) fails, it may be interesting to analyze which dimension induces this failure, namely we can compare the own variance surfaces for the two sources of income targeted by the reform, as well as their covariance surfaces. Such additional tests may be useful for a more in-depth understanding of the impact of marginal program reforms under income variability.

4 Conclusion

The certainty equivalent income of households depends on both expected income and income security. This means that in order to provide stochastic dominance tests for robust poverty comparisons and marginal policy reform orderings, we must deal with income security as well as expected income. The solution is to compare stochastic dominance surfaces rather than curves. The tools could easily be extended to deal with comparisons of social welfare by allowing the maximum poverty lines to exceed the highest levels of certainty equivalent income in the data.

References

Atkinson, A. B. (1987). On the measurement of poverty. *Econometrica, 55*, 759–764.

Bourguignon, F., & Chakravarty, S. R. (2002). *Multi-dimensional poverty orderings*. Paris: Delta, mimeo.

Duclos, J. -Y., Makdissi, P., & Wodon, Q. (2005a). Poverty-dominant transfer programs: The role of targeting and allocation rules. *Journal of Development Economics, 77*, 53–73.

Duclos, J. -Y., Makdissi, P., & Wodon, Q. (2005b). Poverty-reducing tax reforms with heterogeneous agents. *Journal of Public Economic Theory, 7*, 107–116.

Duclos, J. -Y., Makdissi, P., & Wodon, Q. (2008). Socially-improving tax reforms. *International Economic Review, 49*, 1505–1537.

Duclos, J. -Y., Sahn, D., & Younger, S. (2006). Robust multidimensional poverty comparisons. *Economic Journal, 116*, 943–968.

Fishburn, P. C., & Willig, R. D. (1984). Tranfer principles in income redistribution. *Journal of Public Economics, 25*, 323–328.

Foster, J. E., Greer, J., & Torbecke, E., (1984). A class of decomposable poverty measures. *Econometrica, 52*, 761–766.

Foster, J. E., & Shorrocks, A. F. (1988a). Poverty orderings. *Econometrica, 56*, 173–177.

Foster, J. E., & Shorrocks, A. F. (1988b). Poverty orderings and welfare dominance. *Social Choice and Welfare, 5*, 179–198.

Gravel, N., & Tarroux, B. (2015). Robust normative comparisons of socially risky situations. *Social Choice and Welfare, 44*, 257–282.

Makdissi, P., & Wodon, Q. (2002). Consumption dominance curves: Testing for the impact of indirect tax reforms on poverty. *Economics Letters, 75*, 227–235.

Mayshar, J., & Yitzhaki, S. (1995). Dalton improving tax reform. *American Economic Review, 85*, 793–807.

Sen, A. K. (1997). *On economic inequality, expanded edition*. Oxford: Claundon Press.

Yitzhaki, S., & Slemrod, J. (1991). Welfare dominance: An application to commodity taxation. *American Economic Review, 81*, 480–496.

Zheng, B. (1999). On the power of poverty orderings. *Social Choice and Welfare, 3*, 349–371.

Exploring the Covariance Term in the Olley-Pakes Productivity Decomposition

Giannis Karagiannis and Suzanna M. Paleologou

Abstract The aim of this paper is to explore the covariance term in the Olley-Pakes productivity decomposition in order to provide further insights underlying the reallocation effect. In particular, we consider how firms are classified according to their size and productivity scores. We use this information to examine the extent and the importance of the reallocation effect in the Greek cotton industry during a period of high price support. The empirical results show that the policy regime not only allowed least productive farms to survive more than otherwise but also caused the downsizing of the most productive farms and/or the expansion of the least productive ones.

Keywords Olley-Pakes decomposition · Covariance term · Reallocation effect · Price distortions

1 Introduction

The Olley-Pakes (OP) decomposition (Olley and Pakes 1996) is one of the decompositions often used to explain aggregate productivity changes by means of individual firm achievements and resource reallocation. Others decompositions used for similar purposes are those developed by Baily, Hulten and Campbell (1992), Griliches and Regev (1995), Foster, Haltiwanger and Krizan (2001), Baldwin and Gu (2006), Diewert and Fox (2010) and Petrin and Levinsohn (2012). The original formulation of the OP productivity decomposition accounts only for reallocation at the internal margin (i.e., restructuring). However, recent developments have

G. Karagiannis (✉)
Professor, Department of Economics, University of Macedonia, Thessaloniki, Greece
e-mail: karagian@uom.edu.gr

S.M. Paleologou
Associate Professor, Department of Economics, Aristotle University of Thessaloniki, Thessaloniki, Greece
e-mail: smp@econ.auth.gr

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_8

extended it to incorporate the effect of entry and exit (Melitz and Polanec 2015; Maliranta and Maattanen 2015) and to deal with cases where groups of different firms in terms of technology, organization, or ownership co-exist in the same industry (Collard-Wexler and de Loecker 2015).

The reallocation effect in the original formulation of the OP productivity decomposition is depicted by a covariance term between firm size and productivity. In the absence of any market intervention, this term is expected to be positive while in the presence of policy-induced distortions and/or market frictions it may turn zero or even negative. The former is the canonical prediction of firm heterogeneity models (i.e., Jovanovic (1982), Hopenhayn (1992) and Ericson and Pakes (1995)), which presupposes that the largest firms are the most productive while the latter implies that the most productive firms are forced to stay smaller and the least productive to survive and to grow larger than otherwise (see also Restuccia and Rogerson (2008)). According to Bartelsman, Haltiwanger and Scarpetta (2013), the covariance term in the OP productivity decomposition reflects more accurately the meaning of the reallocation effect compared to other dispersion measures used in the literature and thus, it is more relevant for performance evaluation studies as well as for policy analysis.

The aim of this paper is to explore the covariance term (which is equal to the sum across firms of the products of size and productivity deviations from their average values) in the OP productivity decomposition in order to provide further insights for the reallocation effect. For this purpose, we sort out the components of the covariance term into nine groups covering all possible combinations of size and productivity deviations from their average values. The sign and the magnitude of the covariance term depend on the extent of these deviations as well as the percentage of firms in each of these nine groups. Hence, their analysis could provide useful insights of the forces shaping the reallocation effect. For the example, a near zero covariance term may reflect either the absence of almost any productivity and/or size differences across firms or that the combined effect of relatively large firms with above average productivity and of relatively small firms with below average productivity cancels out that of relatively large firms with below average productivity and of relatively small firms with above average productivity. The analytical implications of these two cases are however different as the former points to the representative firm paradigm while the latter to firm heterogeneity models. Moreover, the relative importance of the four groups involving in the latter case is essential for understanding the forces behind the reallocation effect. It is thus important to examine whether or not all of them are involved in determining the covariance term and of course to what extent, as each group has a different impact on aggregate productivity through the reallocation effect. Therefore, from both an analytical and a policy point of view, it really matters how a near zero covariance term emerges in empirical studies. Analogous reasoning applies to the cases of a positive or a negative covariance term.

The rest of this paper is organized as follows: in the next section, we present the OP productivity decomposition and we examine in more details the components of the covariance term. In the third section, we provide background information

about the Greek cotton industry, which consists our case study. Empirical results concerning labor productivity are presented and discussed in the fourth section with particular emphasis on the components of the covariance term and their impact of the reallocation effect. Concluding remarks follow in the last section.

2 The OP Productivity Decomposition

Moving away from the representative firm paradigm, firms are assumed to differ in some observable characteristics such as size, age, capital vintage, etc., as well as in some unobservable characteristics such as entrepreneurial and managerial ability. These differences tend to result in widespread performance heterogeneity with large dispersion and persistence in productivity scores across firms producing the same product. In such a case, each firm could contribute to aggregate (industry-level) productivity growth through its individual performance and its relative importance. Improvements in either one or both would raise aggregate productivity. The firm-level achievements should be aggregated adopting a weighting scheme that accounts for the observed performance heterogeneity. Ideally, the weights used should reflect the relative importance of each unit in the industry as a whole.

Olley and Pakes (1996) proposed the following decomposition to analyze aggregate (industry-level) productivity at a given period t:

$$A_t = \sum_{k=1}^{K} \theta_{it} A_{it} = \overline{A}_t + \sum_{k=1}^{K} \left(\theta_{it} - \overline{\theta}_t\right)\left(A_{it} - \overline{A}_t\right) = \overline{A}_t + \sum_{k=1}^{K} \tilde{\theta}_{it} \tilde{A}_{it} \tag{1}$$

where A refers to a productivity measure and θ to firm size, k is used to index firms, and a bar over a variable denotes its (arithmetic) average value while a tilde over a variable denotes deviations from its average value. Thus, aggregate productivity, which is equal to a weighted average of firm-level productivities with firm size used as weights, is decomposed into two components: an unweighted average of firm-level productivities and a sample covariance between productivity and size. The latter reflects real economic phenomena, namely what has been called market selection mechanism. In an economy driven solely by market forces, selection depends only on market fundaments (i.e., productivity, demand shocks, market power, and input cost). In such a case, as firms located in the low tail of the productivity distribution contract their activity in favor of more productive firms (restructuring), aggregate productivity may increase even if there is no firm-level productivity improvements. For this reason, Bartelsman, Haltiwanger and Scarpetta (2013) argued that the covariance term is an ideal measure to capture the extent of reallocation. Policy distortions and market regulations make the role of productivity and other market fundamentals less relevant for market selection and thus, affect the sign and the magnitude of the covariance term.

The OP decomposition has a clear intuitive interpretation related to the way we may see a number of firms in a group or in an industry as a whole. *First*, we may employ the concept of the representative firm, which implies that there is a firm that can be representative of the whole group or industry or equivalently, the group or industry can be viewed as a replication of the representative firm. In this case, average productivity reflects accurately aggregate productivity. *Second*, we may consider instead the firm heterogeneity models where any firm represents a single point in the productivity and size distributions. In this case, we need a weighted average to reflect aggregate performance accurately, where the productivity of each firm counts but with different weights.[1] The question is then how these two views of the group or the industry differ each other. This depends on whether firms differ in terms of importance and productivity. Apparently, it is essential to account for both and a simple metric that does so is the OP covariance term, which is given by the sum across firms of the products of productivity and size deviations from their averages. Then, aggregate productivity may be seen as the sum of what productivity could have been if all firms were the same (i.e., the representative firm) and the contribution of heterogeneity in terms of both size and productivity that may exist in some industries or groups of firms. The above interpretation questions Balk (2016) intention to call these two sources of aggregate productivity as the Olley-Pakes fallacy.

The covariance term in the OP productivity decomposition is zero if either (a) all firms in the sample or in the industry have the same level of productivity, (b) all firms have the same relative size which would be $\theta_{it} = 1/K$ (for all i), or (c) performance and size are uncorrelated. In these cases, the simple arithmetic average is an unbiased estimate of industry (aggregate) productivity. On the other hand, the covariance term is positive (negative) if either firms with higher than average productivity also have a larger (smaller) than average size and firms with lower than average productivity have a smaller (larger) than average size. Therefore, a positive (negative) covariance term suggests a positive (negative) relationship between size and productivity. Moreover, the larger the covariance term the higher the share of economic activity that goes to more productive firms and thus the higher is aggregate productivity. As a result, the un-weighted average under-estimates (over-estimates) the aggregate industry productivity when the covariance term is positive (negative). The extent of this difference depends on the number of firms for which the covariance term is non-zero and the magnitude of productivity and size deviations from the industry averages.

We may also provide statistical inference about the significance of the two components of the OP productivity decomposition following the econometric model suggested by Hyytinen, Ilmakunnas and Maliranta (2016). They show that the two right-hand side terms in the OP productivity decomposition can be estimated jointly

[1] It is important to determine the right importance metric in the sense that the resulting aggregate productivity measure should have the same meaning and interpretation as the individual measures. We will return to this issue later in the forth section.

by regressing A_{it} on a constant and an appropriately scaled θ_{it} term. Written in terms of a pooled regression equation, this is given as:

$$A_{it} = \alpha_t D_t + \beta_t \theta_{it}^* D_t + \epsilon_{it} \tag{2}$$

where $\alpha_t = \overline{A}_t$, $\beta_t = \sum_{i=1}^{K} \left(\theta_{it} - \overline{\theta}_t\right)\left(A_{it} - \overline{A}_t\right)$, $\theta_{it}^* = \left(\theta_{it} - \overline{\theta}_t\right)/\sigma_t^2 K_t = \left(\theta_{it} - \overline{\theta}_t\right)/\sum_{i=1}^{K}\left(\theta_{it} - \overline{\theta}_t\right)^2$, and σ^2 refers to sample variance, D_t to time dummies, and ϵ_{it} to an *i.i.d.* error term. Based on this regression model we can test the statistical significance of both the average productivity and the covariance term by means of the statistical significance of the estimated parameters α_t and β_t for every time period t.

More useful insights about the covariance term and thus the forces behind the reallocation effect can be gained by classifying its components into nine different groups depending on whether firm size is less, equal or greater than average size and firm productivity is less, equal or greater than average productivity, that is:

$$\begin{aligned}\sum_{k=1}^{K} \tilde{\theta}_{it}\tilde{A}_{it} = &\sum_{\substack{\tilde{\theta}_{it}<0\\ \tilde{A}_{it}<0}} \tilde{\theta}_{it}\tilde{A}_{it} + \sum_{\substack{\tilde{\theta}_{it}<0\\ \tilde{A}_{it}=0}} \tilde{\theta}_{it}\tilde{A}_{it} + \sum_{\substack{\tilde{\theta}_{it}<0\\ \tilde{A}_{it}>0}} \tilde{\theta}_{it}\tilde{A}_{it} + \\ &\sum_{\substack{\tilde{\theta}_{it}=0\\ \tilde{A}_{it}<0}} \tilde{\theta}_{it}\tilde{A}_{it} + \sum_{\substack{\tilde{\theta}_{it}=0\\ \tilde{A}_{it}=0}} \tilde{\theta}_{it}\tilde{A}_{it} + \sum_{\substack{\tilde{\theta}_{it}=0\\ \tilde{A}_{it}>0}} \tilde{\theta}_{it}\tilde{A}_{it} + \\ &\sum_{\substack{\tilde{\theta}_{it}>0\\ \tilde{A}_{it}<0}} \tilde{\theta}_{it}\tilde{A}_{it} + \sum_{\substack{\tilde{\theta}_{it}>0\\ \tilde{A}_{it}=0}} \tilde{\theta}_{it}\tilde{A}_{it} + \sum_{\substack{\tilde{\theta}_{it}>0\\ \tilde{A}_{it}>0}} \tilde{\theta}_{it}\tilde{A}_{it}\end{aligned} \tag{3}$$

From these nine groups, five groups represented by the second, fourth, fifth, sixth and eighth right-hand side terms in (3) make no contribution to the magnitude of the covariance term because either size or productivity or both do not deviate from their averages. From the remaining four groups, two groups represented by the first and the last right-hand side terms in (3) and corresponding to firms with lower than average productivity $\left(A_{it} < \overline{A}_t\right)$ that also have a smaller than average size $\left(\theta_{it} < \overline{\theta}_t\right)$ and to firms with higher than average productivity $\left(A_{it} > \overline{A}_t\right)$ that also have a larger than average size $\left(\theta_{it} > \overline{\theta}_t\right)$ make a positive contribution to the covariance term while the other two groups represented by the third and the seventh right-hand side terms in (3) and corresponding to firms with higher than average productivity $\left(A_{it} > \overline{A}_t\right)$ that also have a smaller than average size $\left(\theta_{it} < \overline{\theta}_t\right)$, and to firms with

lower than average productivity $\left(A_{it} < \overline{A}_t\right)$ that also have a larger than average size $\left(\theta_{it} > \overline{\theta}_t\right)$ make a negative contribution.

From these we can now see that a zero or around zero covariance term may result under two different circumstances: *first*, if all or most of firms fall into the groups with either $A_{it} = \overline{A}_t$ or $\theta_{it} = \overline{\theta}_t$ or both; and *second*, if the sum across firms with $A_{it} < \overline{A}_t$ and $\theta_{it} < \overline{\theta}_t$ and firms with $A_{it} > \overline{A}_t$ and $\theta_{it} > \overline{\theta}_t$ is equal to the sum across firms with $A_{it} < \overline{A}_t$ and $\theta_{it} > \overline{\theta}_t$ and firms with $A_{it} > \overline{A}_t$ and $\theta_{it} < \overline{\theta}_t$. On the other hand, a positive (negative) covariance term may result if the sum across firms with $A_{it} < \overline{A}_t$ and $\theta_{it} < \overline{\theta}_t$ and firms with $A_{it} > \overline{A}_t$ and $\theta_{it} > \overline{\theta}_t$ exceeds (fall short of) the sum across firms with $A_{it} < \overline{A}_t$ and $\theta_{it} > \overline{\theta}_t$ and firms with $A_{it} > \overline{A}_t$ and $\theta_{it} < \overline{\theta}_t$. These sums depend on the magnitude of the productivity and size deviations from the mean and of course on the number of firms in each category. The latter provide information about the group(s) of firms that determine the sign and the magnitude of the covariance term and thus the forces behind the reallocation effect.

We may further examine the role of these nine groups of firms in productivity dispersion. For this purpose, we use the following variance decomposition (see e.g. Juhn et al. 1993):

$$var\left(A_{it}\right) = \sum_{h=1}^{9} m_{ht} var\left(A_{ht}\right) + \sum_{h=1}^{9} m_{ht}\left(\overline{A}_{ht} - \overline{A}_t\right)^2 \tag{4}$$

where h is used to index groups and m refers to their relative size in percentage terms. According to (4), the variance of firm-level productivity in a particular year depends on the variance of productivity within each of the nine groups and the deviations of their group average productivity from the sample mean, each weighted by group's relative size in percentage terms. The first right-hand side term in (4) captures the within-group variance component and the second the between-group component. More widespread performance heterogeneity within a group tends to increase the first component while larger deviations of groups from sample means tend to increase the second component. We will use the variance decomposition analysis to examine whether the extent of performance heterogeneity is related to a negative or a positive contribution to the covariance terms and thus, whether groups more heterogeneous in performance terms tend to increase or not the impact of the reallocation effect.

Next, we provide an illustrative example concerning the cotton industry in Greece during the period that the era of deregulation had started using an unbalanced data set of 1258 farms taken from a harmonized database, the Farm Accounting Data Network (FADN), which collects annual production and cost related data across EU member states.

3 Case Study: Cotton Industry in Greece

Greece is by far the largest supplier of cotton in EU producing around to 3–3.5 times more than Spain, the second larger producer.[2] There is a large number of growers (around 70,000) and cotton accounts for almost 10% of the country's total agricultural output. The main production regions are located in the central and northern parts of the country, where cotton accounts for more than 50% of arable land (Karagiannis 2004). It is grown almost entirely in irrigated land using drip irrigation techniques and is mainly produced by highly specialized farms. Cotton exhibit higher relative profitability compared to competing crops such as maize and durum wheat and for this reason, consist farmers' preferred choice in areas where water supply is sufficient.

The Common Market Organization (CMO) for cotton was introduced in 1981 with Greece's accession to the EU in order to support cotton production and to offer a satisfactory farm income. It is organized on the basis of three policy tools: deficiency payment, co-responsibility levy and maximum quantity guaranteed (MQG). The EU authorities set a target price and its difference from the world price determines the level of the production subsidy that farmers receive at the end of the cropping period. In fact, the deficiency payment is given to cotton ginners on the condition that farmers have received a minimum price per tonne of unginned cotton. The minimum price is associated with the co-responsibility levy mechanism and the MQG. If actual production exceeds the MQG, the minimum and the target prices are reduced by the amount of the co-responsibility levy in order to keep the budgetary cost of aid to predetermined levels.

This intervention scheme remained mainly unchanged until the most recent CAP reform when it was transformed into a scheme with direct income and production aid (i.e., decoupling). Whatever changes in the CMO for cotton prior to 2005 regarded the way MQG and the co-responsibility levy were determined. In the period 1981–1995, the MQG was set at the EU level while in the period 1996–2005, it was set at a national level for each member-state. Accordingly, the level of the co-responsibility levy was determined per fixed amounts of excess production during the period 1981–1995 and as a percentage of excess production since then (Karagiannis 2004). Notice that in the period prior to Spain and Portugal accession to the EU, the scheme operated as a pure deficiency payment as the MQG was never exceeded. The co-responsibility levy mechanism was activated afterwards as Community production exceeded the MQG.

Our analysis is focused on the last part of the period when the MQG was set at the EU level, namely the period 1991–1995. During this period, Community production steadily exceeded the MQG and Spanish growers felt that the uniform reduction of the target price throughout the EU was unfair as the fast production expansion in Greece was responsible for the violation of the MQG. Political pressure from

[2]In ginned cotton, the EU is considered as a small trading country accounting for around 5% of world imports. A more detailed discussion of the world cotton market see Baffes (2004).

their side induced afterwards the division of the Community MQG into national quantities guaranteed. In the case that government agencies adjust policy regimes in response to industry performance levels and/or industry lobbying, as seems to be the case with the introduction of national quantities guaranteed, this may induce a spurious correlation between productivity and policy changes. By restricting attention to the period immediately before or after the policy changes we can, according to Topalova and Khandelwal (2011), mitigate the confounding effect that may arise because of possible endogeneity of policy changes.

Distortions and market interventions tend in general to squeeze the importance of market forces in shaping industry-level productivity and thus decrease the relative importance of the reallocation effect (see Retruccia and Rogerson 2008; Hsieh and Klenow 2009; Eslava et al. 2013). Price support schemes in particular lower the level of productivity needed to earn positive expected profits and this makes it a lot easier for the relatively low productivity firms to survive. This implies a higher concentration of industry output resulting from firms in the lower tail of productivity distribution and it is consistent with a weaker market selection mechanism (Aw et al. 2003). In addition, policy distortions may prevent efficient firms from achieving optimal scale, increase managerial slack, shrink innovation adoption incentives and keep inefficient firms from contracting economic activity (Hsieh and Klenow 2009). All these according to Bartelsman, Haltiwanger and Scarpetta (2009) tend to decrease the magnitude of the covariance term in the OP decomposition and even turn it negative, as in de Loecker and Konings (2006). As a result, the difference between the weighted and the unweighted average of individual productivities would be small.

The data for the present study are from a harmonized European database, the Farm Accounting Data Network (FADN). The FADN provides annual statistics on the state of agriculture in the EU based on a sample of almost 60,000 farms, around 10% of which are located in Greece. Data are collected from a rotating panel of farms. The FADN field of observation covers large entrepreneurial farms as defined in the farm structure survey of the EU and excludes smaller farms below FADN thresholds; in particular, farms with less than 2 ESU (European Size Units) are excluded. The relevant data are collected in a consistent manner across EU member countries using the same methodology and accounting standards.

In the FADN, farms are classified by commodity according to their source of revenue. That is, a farm is classified as cotton-oriented as long as at least two thirds of its revenue come from cotton production. During the period under consideration, 2827 observations were available for cotton producers in Greece. These referred to a total of 1258 farms implying that on average each farm is observed 2 to 3 times during the 5 years period. In particular, the sample includes 436 farms in 1991, 526 in 1992, 553 in 1993, 600 in 1994, and 722 in 1995.

4 Empirical Results

In the present study, we use labor productivity as our preferred performance measure. Output is measured in terms of deflated total gross farm revenue and labor, which includes both family and hired workers, is measured in annual working hours. The gross revenue of each farm is deflated by a common industry price index. The use of a common industry deflator is not expected to induce any measurement errors in our case as the cotton industry contains no elements of horizontal product differentiation and any output price variation reflects quality differences.

The second issue related to the estimates of labor productivity reported in Table 1 is the choice of aggregation weights. Consistency in aggregation requires the resulting aggregate labor productivity measure to have the same interpretation as the firm-level labor productivity measures. Following van Biesebroeck (2008) and Färe and Karagiannis (2017), this is ensured as long as the weights are defined in terms of the variable that is in the denominator of the ratio-type productivity measure. For labor productivity, this implies that the aggregation weights should be in terms of the labor input rather than in terms of output market share as is often done in the literature. Only in this case the aggregate productivity measure reflects the ratio of total industry or group output to total industry or group labor input.

The empirical results for the OP decomposition of labor productivity are given in Table 1. From there we can see that the aggregate productivity was lower than average productivity for the whole period under consideration. The negative sign of the covariance term reflects the distortion placed in the unginned cotton market by means of a target price, the MQG, and the associated co-responsibility levy. Such policy measures often tend to squeeze the relative importance of the reallocation term because they allow unproductive farms to stay in business more than otherwise and as a result, a relatively smaller market share remains for reallocation. However, there is a declining trend in the size of the covariance term as the changes in policy regime induced some competitive pressure to farmers. In relative terms though the importance of individual achievements, as reflected in the average productivity, were more important in shaping aggregate productivity than the reallocation effect.

The estimated parameters of the regression model reported in Table 2 confirm the statistical importance of both the average productivity and the reallocation effect for

Table 1 Decomposition of aggregate labor productivity in Greek cotton industry

	Aggregate labor productivity	Average labor productivity		Covariance term	
	(1)	(2)	(3)	(4)	(5)
1991	3.726	4.423	*118.7*	−0.698	*−18.7*
1992	3.145	3.672	*116.8*	−0.527	*−16.8*
1993	3.958	4.445	*112.4*	−0.490	*−12.4*
1994	4.010	4.595	*114.6*	−0.585	*−14.6*
1995	4.028	4.395	*109.1*	−0.367	*−9.1*

Note: Columns (3) and (5) give the percentage contribution of average productivity and the covariance term in aggregate productivity.

Table 2 Statistical inference of the aggregate labor productivity decomposition components for the Greek cotton industry

Parameters	Estimated value	*t*-statistic	95% Confidence interval
α_{1991}	4.423	23.45	[4.053, 4793]
α_{1992}	3.672	34.35	[3.462, 3.881]
α_{1993}	4.445	39.03	[4.222, 4.669]
α_{1994}	4.595	43.51	[4.388, 4.802]
α_{1995}	4.395	49.44	[4.221, 4.570]
β_{1991}	−0.698	−4.84	[−0.980, −0.415]
β_{1992}	−0.527	−10.23	[−0.628, −0.426]
β_{1993}	−0.490	−9.40	[−0.592, −0.387]
β_{1994}	−0.585	−10.13	[−0.698, −0.472]
β_{1995}	−0.367	−7.89	[−0.459, −0.276]

all years from 1991 to 1995. In addition, the individual statistical significance of the covariance terms implies that average productivity cannot represent accurately the aggregate productivity of the industry in any of the years under consideration.[3] Consequently, the representative firm paradigm is not supported in our case. Instead, the empirical results suggest that farms heterogeneity and reallocation at the internal margin (i.e., restructuring) have contributed negatively to aggregate productivity in a statistically significant way. This provides the basis for further analysis of the covariance term to examine the forces behind restructuring in the Greek cotton industry.

The relevant results are reported in Table 3 and from there we can see that, on average over the period under consideration, we have 39.3% of the farms with a positive contribution to the covariance term, 57.1% with a negative contribution and only 3.6% with no contribution.[4] These figures provide a first interpretation for the negative sign of the covariance term. A further look in Table 3 indicates that the vast majority of farms (around 86%) is almost equally split into three groups, namely the relatively small farms with below average labor productivity (28.5%), the relatively small farms with above average labor productivity (28.8%), and the relatively large farms with below average labor productivity (28.3%). Of these three groups, the latter two contribute negatively to the covariance term and the former positively. Even if we account for the 10.8% of farms in the group of relatively large farms with above average productivity, which also contribute positively, the covariance term ends up negative.

[3] We also reject at any level of significance the hypothesis that the covariance terms for all periods are jointly equal to zero as the value of the calculated $F(5, 2817) = 59.00$ is larger than the tabulated one. In addition, we reject at the 5% level of significance the hypothesis that the covariance term was common to all periods as the value of the calculated $F(4, 2817) = 3.21$ is larger than tabulated one.

[4] Notice that we found no farms with average productivity and for this reason Table 3 contains six instead of nine entries for each year.

Table 3 Analysis of the covariance term for the Greek cotton industry

		$A_{it} - \overline{A}_t < 0$		$A_{it} - \overline{A}_t > 0$	
$\theta_{it} - \overline{\theta}_t < 0$	1991	0.160	*27.5*	−0.473	*30.0*
	1992	0.141	*26.8*	−0.358	*30.4*
	1993	0.175	*30.9*	−0.357	*27.7*
	1994	0.155	*27.8*	−0.381	*29.3*
	1995	0.149	*29.8*	−0.314	*26.7*
$\theta_{it} - \overline{\theta}_t > 0$	1991	0	*1.9*	0	*0.2*
	1992	0	*1.7*	0	*1.1*
	1993	0	*2.0*	0	*1.8*
	1994	0	*2.3*	0	*1.7*
	1995	0	*2.5*	0	*2.2*
$\theta_{it} - \overline{\theta}_t = 0$	1991	−0.430	*31.5*	0.045	*8.9*
	1992	−0.358	*29.5*	0.049	*10.5*
	1993	−0.369	*27.3*	0.062	*10.3*
	1994	−0.431	*27.8*	0.072	*10.5*
	1995	−0.292	*25.5*	0.090	*13.3*

Notes: (1) The sum of the nine entries per year gives the covariance terms reported in column (4) of Table 1.
(2) In italics are the percentages of farms in each group per year.

In terms of farm numbers, among the relatively small ones (which on average account for 57%), the percentage of those with above average productivity is on average equal to the percentage of those with below average productivity. In contrast, among the relatively large farms (which on average account for 40%), the percentage of those with below average productivity is on average greater than the percentage of those with above average labor productivity. In particular, 75% of relatively large farms achieved below average productivity and the remaining 25% above average productivity. Therefore, from a size perspective, the group of relatively large farms with below average labor productivity is the one that essentially determines the negative sign of the covariance term. Alternatively, among the farms with below average productivity (which on average account for 58%), the percentage of small farms is on average equal to the percentage of large farms while among the farms with above average productivity (which on average account for 42%) the percentage of small farms is on average greater than that of large farms. In particular, 75% of farms with above average productivity are relatively small and the remaining 25% are relatively large. Therefore, from a performance perspective, the category of relatively small farms with above average productivity is the one that determines the negative sign of the covariance term.

Nevertheless, there was no a one-to-one correspondence between the number of farms in each group and contribution of each group into the covariance term. That is, smaller deviations, particularly in terms of labor productivity, were found for the two groups with positive covariance components compared to those for the two groups with negative covariance components. These two latter groups made an equal contribution to the covariance term, which on average was around 0.376, while the

contribution of the group of the relative small farms with below average productivity was much smaller; in fact, less than half (0.156) of the other two. The contribution of the other group with positive covariance components, namely that of the relatively large farms with above average productivity, was even smaller, around 0.064.

During the period under consideration, the number of farms in the two groups with negative covariance components declined while that in the two groups with positive covariance components increased, reflecting the competitive pressure from policy changes. For this reason, the magnitude of the covariance term decreased over time. On the other hand, the contribution of relatively small farms in the covariance term fell while that of relatively large farms rose over time. The latter is mainly due to the gradual increase of the contribution of the relative large farms with above average productivity, which doubled from 0.045 in 1991 to 0.090 in 1995 (see Table 3). This indicates another positive aspect of the accomplished policy changes that enhance the role of relatively large farms with above average labor productivity in shaping the reallocation effect in the Greek cotton industry.

The empirical results related to variance decomposition are reported in Table 4. From there we can see that the groups with above average productivity exhibited larger performance heterogeneity, as reflected in their higher within-group variances, compared to the groups with below average productivity. For the whole period, the group of relatively small farms with above average productivity had by far the larger performance heterogeneity, followed by the group of relatively large farms with below average productivity. These are the two groups with negative covariance components. From Table 4 we can see that the contribution to productivity dispersion of the groups with negative covariance components is greater than that of the groups with positive covariance components. Furthermore,

Table 4 Variance decomposition of labor productivity in Greek cotton industry

Group		1991	1992	1993	1994	1995
$\theta_{it} - \overline{\theta}_t < 0$	$\overline{A}_{ht}$	2.617	2.088	2.729	2.957	2.136
$A_{it} - \overline{A}_t < 0$	var(A_{ht})	0.959	0.809	1.005	0.850	0.668
$\theta_{it} - \overline{\theta}_t < 0$	$\overline{A}_{ht}$	7.712	6.353	7.441	7.303	6.832
$A_{it} - \overline{A}_t > 0$	var(A_{ht})	31.982	6.197	5.638	4.871	5.878
$\theta_{it} - \overline{\theta}_t = 0$	$\overline{A}_{ht}$	2.397	1.919	2.380	2.947	2.938
$A_{it} - \overline{A}_t < 0$	var(A_{ht})	1.325	0.323	0.657	0.705	1.239
$\theta_{it} - \overline{\theta}_t = 0$	$\overline{A}_{ht}$	5.294	5.276	8.523	6.327	6.614
$A_{it} - \overline{A}_t < 0$	var(A_{ht})	0	3.423	8.571	2.085	5.581
$\theta_{it} - \overline{\theta}_t > 0$	$\overline{A}_{ht}$	2.412	1.857	2.591	2.568	2.653
$A_{it} - \overline{A}_t < 0$	var(A_{ht})	0.809	0.724	1.221	1.068	0.928
$\theta_{it} - \overline{\theta}_t > 0$	$\overline{A}_{ht}$	6.401	5.159	6.144	6.528	6.012
$A_{it} - \overline{A}_t > 0$	var(A_{ht})	3.555	1.340	2.082	4.411	2.056
	var(A_{it})	16.447	6.630	7.581	7.167	5.903
	within-group	10.454	2.498	2.585	2.474	2.448
	between-group	5.843	4.141	5.015	4.670	3.502

Note: The m_{ht} values for the calculation of the within-group and the between-group components are given in Table 3 in the column with italics.

the three groups that contribute the most to the reallocation effect are the same ones that contribute the most to performance heterogeneity. On the other hand, in all but the first sample year, the within-group component dominated the between-group component indicating that group differences from average productivity were more important in determining productivity dispersion than within-group variability.

5 Concluding Remarks

In this paper we explore the covariance term in the OP productivity decomposition in order to examine the extent and the importance of the reallocation effect in the Greek cotton industry. We first examine whether and to what extent the reallocation of economic activity towards more productive farms and/or away from less productive ones has been productivity enhancing. The empirical results clearly illustrate that it did not. Due to policy-induced distortions, it had a negative effect on aggregate productivity. The then CAP support system not only allowed least productive farms to survive more than otherwise but also allowed the downsizing of the most productive farms and/or the expansion of the least productive ones. The detailed analysis of the covariance term shown that, from a performance perspective, the category of relatively small farms with above average productivity is the one that mainly determines the negative sign of the covariance term while, from a size perspective, the category of relatively large farms with below average labor productivity is the main responsible.

References

Aw, B. Y., Chung, S., & Roberts, M. J. (2003). Productivity, output and failure: A comparison of Taiwanese and Korean manufacturers. *Economic Journal, 113*, F485–F510.

Baldwin, J. R., & Gu, W. (2006). Plant turnover and productivity growth in canadian manufacturing. *Industrial Corporation and Change, 15*, 417–465.

Balk, B. M. (2016). The dynamics of productivity change: A review of the bottom-up approach. In W. H. Greene (Ed.), *Productivity and efficiency analysis* (pp. 15–49). Switzerland: Springer Int. Publ.

Baffes, J. (2004). Cotton: Market setting, Trade policies and issues, *World Bank Policy Research Working Paper* 3218.

Baily, M.N., Hulten, C.R. and D. Campbell (1992). Productivity dynamics in manufacturing plants, *Brookings Papers on Economic Activity*, 187–267.

Bartelsman, E., Haltiwnger, J., & Scarpetta, S. (2009). Measuring and analyzing cross-country differences in firm dynamics. In T. Dunne, B. J. Jensen, & M. J. Roberts (Eds.), *Producers ynamics: New evidence from micro data, NBER, studies in income and wealth* (Vol. 68, pp. 15–76). Chicago: University of Chicago Press.

Bartelsman, E., Haltiwanger, J., & Scarpetta, S. (2013). Cross-country differences in productivity: The role of allocation and selection. *American Economic Review, 103*, 305–334.

van Biesebroeck, J. (2008). Aggregating and decomposing productivity. *Review of Business and Economics, 2*, 122–146.

Collard-Wexler, A., & de Loecker, J. (2015). Reallocation and technology: Evidence from the US steel industry. *American Economic Review, 105*, 131–171.

Diewert, W. E., & Fox, K. J. (2010). On measuring the contribution of entering and exiting firms to aggregate productivity growth. In W. E. Diewert et al. (Eds.), *Price and productivity measurement. Index number theory* (Vol. 6). Victoria: Trafford Press.

Ericson, R., & Pakes, A. (1995). Markov-perfect industry dynamics: A framework for empirical work. *Review of Economic Studies, 62*, 53–82.

Eslava, M., Haltiwanger, J., Kugler, A., & Kugler, M. (2013). Trade and market selection: Evidence from manufacturing plants in Colombia. *Journal of Economic Dynamics, 16*, 135–158.

Färe, R., & Karagiannis, G. (2017). The denominator rule for share-weighting aggregation. *European Journal of Operational Research, 260*, 1175–1180.

Foster, L., Haltiwanger, J., & Krizan, C. J. (2001). Aggregate productivity growth: Lessons from microeconomic evidence. In C. R. Hulten, E. R. Dean, & M. J. Harper (Eds.), *New developments in productivity analysis, NBER, studies in income and wealth* (Vol. 63, pp. 303–363). University of Chicago Press.

Griliches, Z., & Regev, H. (1995). Firm productivity in Israeli industry, 1979-1988. *Journal of Econometrics, 65*, 174–203.

Hopenhayn, H. A. (1992). Entry, exit and firm dynamics in long-run equilibrium. *Econometrica, 60*, 1127–1150.

Hsieh, C. T., & Klenow, P. J. (2009). Misallocation and manufacturing TFP in China and India. *Quarterly Journal of Economics, 124*, 1403–1448.

Hyytinen, A., Ilmakunnas, P., & M. Maliranta. Olley-Pakes productivity decomposition: Computation and inference, *Journal of Royal Statistical Society Series A*, 2016 (forthcoming).

Jovanovic, B. (1982). Selection and evolution of industry. *Econometrica, 50*, 25–43.

Juhn, C., Murhpy, K. M., & Pierce, B. (1993). Wage inequality and the rise in returns to skill. *Journal of Political Economy, 101*, 410–442.

Karagiannis, G. (2004). The EU cotton policy regime and the implications of the proposed changes for producer welfare, *FAO Commodity and Trade Policy Research Working Paper No 9*, Rome.

de Loecker, J., & Konings, J. (2006). Job reallocation and productivity growth in a post-socialist economy: Evidence from Slovenian manufacturing. *European Journal of Political Economy, 22*, 388–408.

Maliranta, M., & Maattanen, N. (2015). An augmented Olley-Pakes productivity decomposition with entry and exit: Measurement and interpretation. *Economica, 82*, 1372–1416.

Melitz, M. J., & Polanec, S. (2015). Dynamic Olley-Pakes productivity decomposition with entry and exit. *RAND Journal of Economics, 46*, 362–375.

Olley, G. S., & Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica, 64*, 1263–1297.

Petrin, A., & Levinsohn, J. (2012). Measuring aggregate productivity growth using plant level data. *RAND Journal of Economics, 43*, 705–725.

Restuccia, D., & Rogerson, R. (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic Dynamics, 11*, 707–720.

Topalova, P., & Khandelwal, A. (2011). Trade liberalization and firm productivity: The case of India. *The Review of Economics and Statistics, 93*, 995–1009.

The Decline of Manufacturing in Canada: Resource Curse, Productivity Malaise or Natural Evolution?

Robert Petrunia and Livio Di Matteo

Abstract The state of Canadian manufacturing is a constant issue in current economic and public policy debates. Over the past 50 years, there has been a decline in the contribution of manufacturing to the overall Canadian economy. This decline is especially true for the economies of two provinces, Ontario and Quebec, which traditionally were the most manufacturing intensive. Ontario, in particular, is hit hard by the 2008 recession in terms of both employment and output share decline of its manufacturing sector. This paper explores the relative importance of three explanations for the decline of the Canadian manufacturing sector. Natural evolution offers the first explanation as the economies of most Western countries move away from manufacturing and toward the services. A second explanation for this decline is Dutch Disease. The period from 2003 to 2014 sees both a significant rise in commodity prices and the Canadian dollar. This period also saw a booming resource sector – in particular, the energy / commodity producing provinces of Alberta, Saskatchewan and Newfoundland and Labrador. Finally, Canada's manufacturing productivity performance has been weak relative to other countries, which may also be a factor in its manufacturing decline. Our results show that most of the Canadian manufacturing sector decline occurs in Ontario and Quebec, while manufacturing's contribution remains flat or slightly increases in most of the other provinces.

Keywords Manufacturing · Resource curse · Productivity

JEL Classification: D2, L1, L6, N5, N6, R1

R. Petrunia (✉) • L. Di Matteo
Lakehead University, Thunder Bay, ON, Canada
e-mail: rpetruni@lakeheadu.ca; ldimatte@lakeheadu.ca

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_9

1 Introduction

The state of Canadian manufacturing is a constant issue in current economic and public policy debates especially with respect to the economies of central Canada. Both Ontario and Quebec have seen manufacturing employment decline with Ontario particularly hard hit during the 2009 recession. Between 2000 and 2013, manufacturing employment in Ontario drops from 1,072,000 jobs to 777,300 jobs – a drop of 27% – while in Quebec it drops from 629,000 to 486,000 – a 23% decline. In 2000, Ontario and Quebec together accounted for 76% of Canada's 2,242,300 manufacturing jobs whereas by 2013 it was 73% of 1,734,200 jobs.[1]

The explanations for this manufacturing decline and the lament of deindustrialization have often focused on a Dutch disease or resource curse explanation. A booming resource sector – in particular, energy in Alberta, Saskatchewan and Newfoundland and Labrador- ostensibly contribute to the Canadian dollar's appreciation which in turn causes manufacturing exports to decline.[2] The relationship between resources and a shrinking manufacturing sector in response to exchange rate and productivity effects caused by a booming resource sector was explored by the work of Corden and Neary (1982, 1983) during the North Sea oil boom. Furthermore, Sachs and Warner (1995, 1999, 2001) show resource abundant economies grew slower than resource scarce economies since 1970 and dubbed this phenomenon the "Curse of Resources."

Resource exports have always been important in Canadian economic history but manufacturing decline has generated some discomfort with our resource sector and its alleged impact on the rest of the economy. This is despite the fact that nearly half of Canada's total manufacturing output – from pulp mills to automobile production – is indeed still resource based.[3] The resource sector is also vital to the economic health of our transportation sector as natural resource products account for more than two-thirds of rail and marine shipments in Canada. This suggests that

[1]Data Source: Table 2 820012 - Labour force survey estimates (LFS), employment by class of worker, North American Industry Classification System (NAICS) and sex, annually (Persons).

[2]Natural resources have been an important driver of general Canadian economic prosperity. For Canada, Keay (2007) finds that the exploitation of Canada's natural resources during the 20th century made direct and indirect contributions to the size and efficiency of the Canadian economy and had a substantial positive impact on the level of real per capita GDP, contributing about 20 percent. Another comprehensive study by Baldwin and MacDonald (2012) also finds natural resources and trade to be important contributors to Canadian real gross national income between 1870 and 2010.

[3]Cross (2015) states that 46.2 percent of all manufacturing output in Canada in 2010 was resource based.

there are substantial economic linkages between Canada's resource sector and the rest of the economy.[4]

Other explanations of this manufacturing decline have sought other sources of changes in Canada's manufacturing sector. Cross (2013), for example, argues that the appreciation of the Canadian dollar in the early twenty-first century was not driven solely by commodity prices but was due to a decline in the value of the US dollar and increased investment flows into Canada. Moreover, the struggles in Canadian manufacturing affected mainly automobile production, clothing and forestry related manufacturing and these industries also all contracted in the United States at the same time. Baldwin and MacDonald (2009) actually find little evidence of long-term manufacturing decline as in terms of volume of output, Canadian manufacturing production as a share of the economy has not changed much in about half a century.

Capeluck (2015a, b) finds that demand side factors and outsourcing have been factors explaining the decline in Canadian manufacturing but labour productivity growth has been a particularly important factor. Capeluck finds above average labour productivity growth explains most of the decline in the manufacturing employment share before 2000, while the post 2000 decline is explained by a loss in cost competitiveness linked to an appreciation of the Canadian dollar; increased competition in the U.S. import market; and a slowdown in domestic demand growth in the United States.

This paper provides an examination of three explanations for Canada's manufacturing decline. First, a comparison is done with other developed countries to see if Canada differs markedly from them or is also part of a gradual evolution away from goods production underway in most developed countries. Second, is the decline of manufacturing simply a function of Canadian productivity growth being weaker than other countries? Third, can Canada's manufacturing decline be attributed to a resource curse argument related to the appreciation of Canada's currency especially with regards to its major trade partner-the United States. The overall results of our analysis suggest that Canada's manufacturing decline is more related to general international trends in manufacturing decline and productivity results rather than the effects of currency appreciation.

[4]"Staples" (natural resources for export) approaches to economic development describe a process by which "linkages" associated with the natural resource production encourage industrialization provided the linkages are strong enough, and the income associated with them is retained in the domestic economy. The economic development of resource abundant, sparsely populated regions has been explained by the classic staples approach or models of export-led development as originally set out in the work of H.A. Innis (1969[1956]) who followed earlier work by G.S. Callender (1902, 1965[1909]) and W.A. Mackintosh (1923). The classic works on Canadian staples by Innis are The Fur Trade in Canada (1984[1930]) and The Cod Fisheries (1978[1940]). Modern versions of staple theory see economic development as a process of diversification around an export base. For relevant literature, see the papers by Baldwin (1956), Watkins (1963) and Caves (1966, 1971).

2 Decline of Manufacturing Across Canada and the G7 Countries

Canada has witnessed a decline in manufacturing's share of GDP since 1926 (See Fig. 1). From 1926 to 1943, the manufacturing share of GDP rose from 20% to 29% and then declined to reach 18% by 1981. Between 1981 and 2001, the manufacturing to GDP ratio stabilized averaging 17% and ranging from a high of 19% to a low of 16%. Since 2001, the ratio has declined going from 19% to approximately 13% by 2014.

This manufacturing decline is a feature of numerous other economies around the world. Figures 2a–c and 3 a, b present United Nations data on the manufacturing share of total value-added in the other G-7 and in world regions from 1970 to 2013. Between 1970 and 2014, the manufacturing sector's share of total value added declined in all of the G-7 countries with the exception of Japan and Italy where it rose slightly. The average percentage point decline across the G-7 countries between 1970 and 2014 was 2.6% with the largest declines registered by Germany and the UK at −6.5 and −8.8% points respectively with Canada as the next largest drop at −3.7 points. Compared to the G-7 as a whole, Canada has generally always had a lower manufacturing to GDP ratio. The G-7 as a whole sees a continuous decline in the manufacturing share of GDP over the period, while Canada sees a sharp decline since 2001. However, Fig. 2c indicates a strong positive correlation between the two.

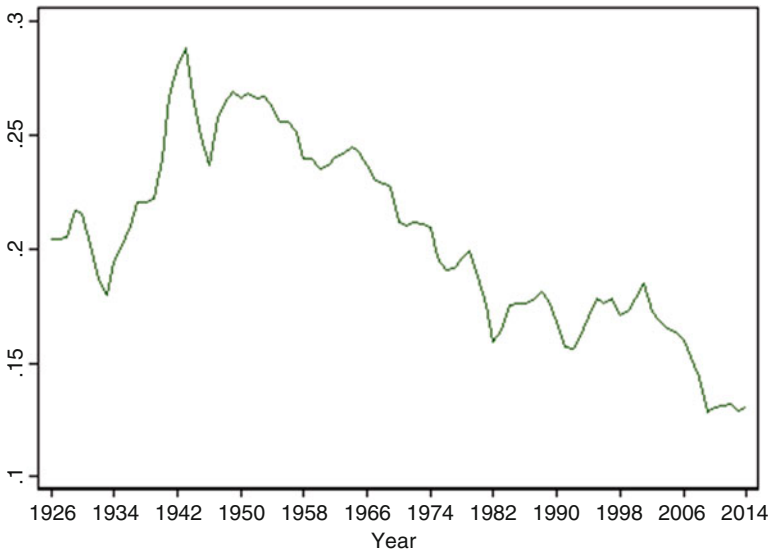


Fig. 1 Manufacturing to GDP Ratio, Canada, 1926–2014

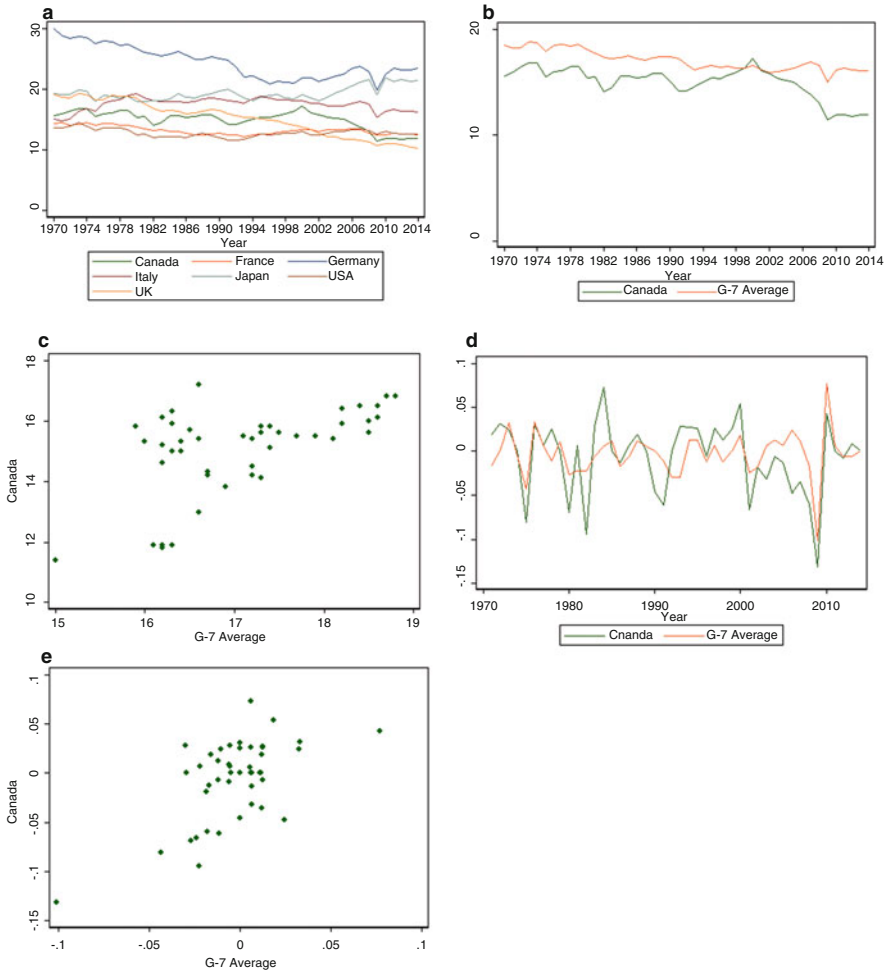


Fig. 2 (**a**) Manufacturing value added as a percent share of total value added,1970 to 2014, G-7 countries (**b**) Manufacturing value added as a percent share of total value added, 1970 to 2014, Canada and G-7 countries minus Canada Average. (**c**) Comparison of manufacturing value added as a percent Share of Total Value Added,1970 to 2014, Canada and G-7 countries minus Canada average (**d**) Growth of manufacturing percent share of total value added,1970 to 2014, Canada and G-7 countries minus Canada average (**e**) Comparison of growth of manufacturing percent share of total value added,1970 to 2014, Canada and G-7 countries minus Canada average

To better capture short-term fluctuations, Fig. 2d, we compare the growth rate of the manufacturing sector's share of value added for both Canada and the average for the other G-7 countries.[5] The movement of this growth rate for Canada and the other

[5]The growth rate is calculated as the difference in the logarithm values of the variables.

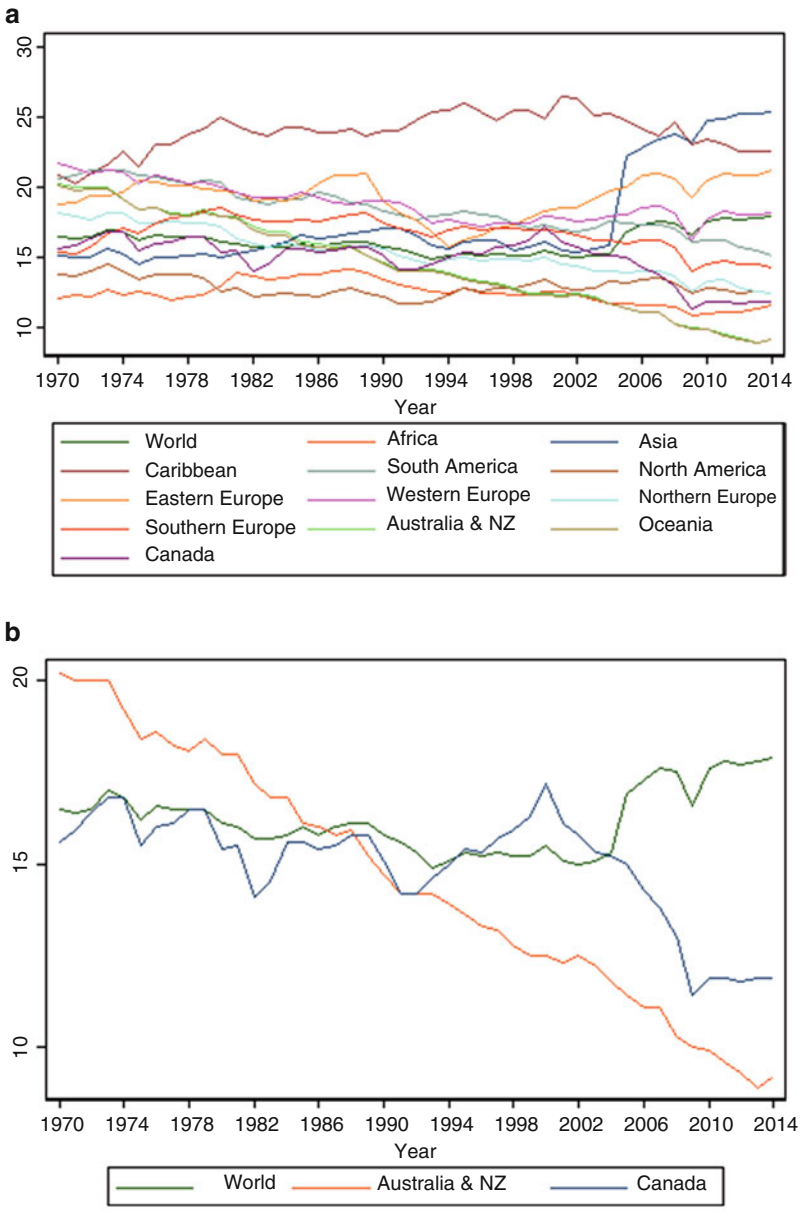


Fig. 3 (**a**) Manufacturing value added as a percent share of total value added World regions, 1970–2014 (**b**) Manufacturing value added as a percent share of total value added Canada, Australia & New Zealand and World, 1970–2014

Table 1 Regressions comparing manufacturing sector's share of value added, Canada and G-7 minus Canada average (1970–2014)

	Dependent Variable	
	Canada Manu Share	Growth Canada Manu Share
G-7 Manu Share	0.937***(0.199)	–
Growth G-7 Manu Share	–	0.991***(0.198)
Constant	−1.014(3.402)	−0.003(0.005)

Note: ***, **, and * indicate significance at the 1% level, 5% level and 10% level, respectively. Standard errors are in parentheses

G-7 countries follow similar patterns over the 1970 to 2013 period. Prior to 2000, the manufacturing sector's share growth experiences both positive and negative years but there is a gradual drop in the manufacturing sector's contribution due to the larger absolute value during negative years. The post 2000 period shows mainly negative growth in manufacturing's share of the Canadian economy. Both Canada and the G-7 countries experience a large drop in their manufacturing sector's share in 2008, but this substantial drop is followed by a large rise in 2010.

To summarize the relationship, Table 1 provides results for a regression of the manufacturing share of GDP in Canada against the average in the other G-7 countries. Column 1 presents estimates using the levels variables, while column 2 shows estimates comparing growth rates. The results indicate that Canada's manufacturing share of GDP moves almost in tandem with the other G-7 countries both in terms of levels and growth rates. In both cases, the coefficient on the G-7 variable is not statistically significantly different from one. A general picture emerges from Table 1 and Fig. 2a–e. The long-term trend is that the contribution of manufacturing to the economy is declining across all the G-7 countries. Further, short-term fluctuations in this contribution are also similar across the G-7 countries.

Over the period 1970 to 2014, the manufacturing share of world economic activity has remained relatively stable ranging from 15% to 18%. Figure 3a, b look at manufacturing's contribution in various world regions. However, with the exception of Asia, the Caribbean, and Eastern Europe, most regions have seen a decline in the manufacturing share of economic activity. When compared to the world as a whole, Canada's manufacturing share of economic activity has been close to the world average until approximately 2003 when it has fallen dramatically below. However, Canada has still managed to retain more manufacturing than Australia and New Zealand – countries with a resource export orientation that parallels Canada. Indeed, over the period 1970 to 2014, Canada saw its manufacturing share of valued added fall from 17% to 11% while Australia and New Zealand combined fell from 20% to 9%.

Any discussion of manufacturing decline in Canada invariably focuses on the sector's declining share of GDP and employment. In the case of employment, there is a difference in performance between the sector share of total employment – which has seen a decline – and the absolute level of employment which has fluctuated but at present is not that much different than it was several decades ago. Figure 4

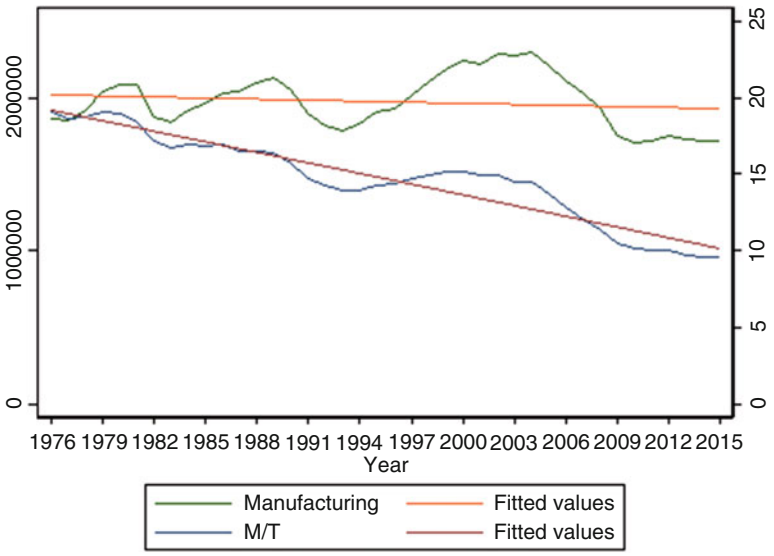


Fig. 4 Manufacturing employment in Canada total and employment shares, 1976–2015

illustrates the sector's employment performance in terms of total job numbers and total employment share from 1976 to 2015. Total manufacturing employment has been marked by several cycles of increase and decrease with the decline period since 2000 amongst the steepest. Yet, total manufacturing employment in 2015 is only 8% lower than 1976. On the other hand, manufacturing's share of total employment has been much more pronounced dropping by 50% over the same period.

3 Productivity Decline

Given the growth in value of manufacturing output over time, another important variable is productivity. Figure 5 presents Canadian manufacturing value added per employee in 2005 constant U.S. dollars over the period 1970 to 2008 using United Nations data and compared to a select number of other countries.[6] The increase in Canadian manufacturing value added per employee is striking. However, Canadian productivity growth in manufacturing has slowed since 2000 especially compared to the United States and Japan – both of which have not seen the decline in manufacturing seen by Canada. Indeed, of these five countries over the period 2000 to 2008, Canada actually had the second lowest average annual growth rate in

[6]Huynh et al. (2011) and Huynh and Petrunia (2016) examine labor productivity growth for firms in the manufacturing sector.

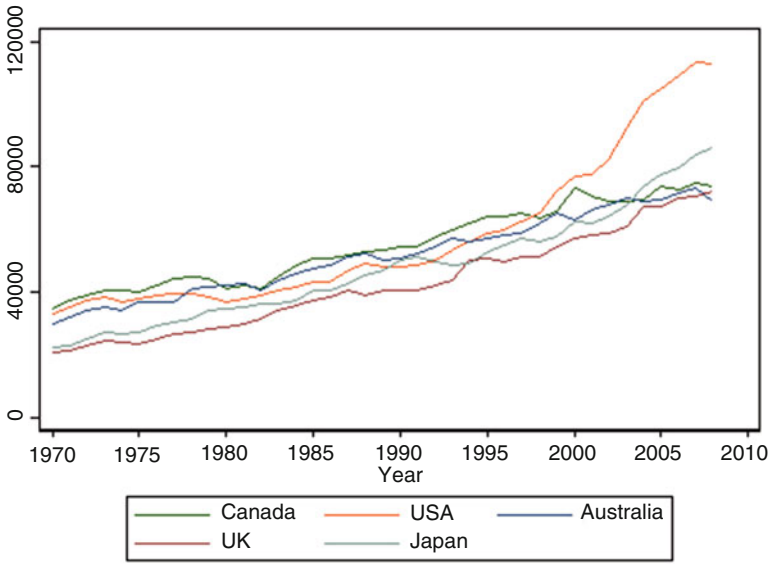


Fig. 5 Manufacturing value added per employee, 1970 to 2008

manufacturing value added per employee at 1.4%. The United States and the United Kingdom in contrast had growth rates of 5.2% and 3.3% respectively while Japan's was 4.5%. Only Australia was lower at 0.7%.

Figure 6a–d provide a comparison of the manufacturing sector's share of Canadian GDP to labour productivity of the Canadian manufacturing sector. Labour productivity measures real output per employee. Figure 6 demonstrates a negative relationship between the share of manufacturing and labour productivity in manufacturing. Figure 6b confirms this relationship by looking at the time series patterns of the two variables. Naturally, labour productivity in the Canadian manufacturing factor increases over time with the accumulation of capital and technological growth.

One problem with comparing levels of manufacturing sector's share of GDP with labour productivity is the relationship may be spurious due to the time series properties of the variables. To address any potential spurious correlation and look at the short term fluctuations, Fig. 6c, d compare the growth of the two variables. These figures indicate a positive short term growth relationship between the two variables. Further, Table 2 provides the results from the regression of the growth rate of manufacturing's share on its labour productivity growth and confirms that there is a strong positive unconditional relationship between these two variables over the short run. Thus, declining labour productivity growth can be considered a factor in manufacturing decline.

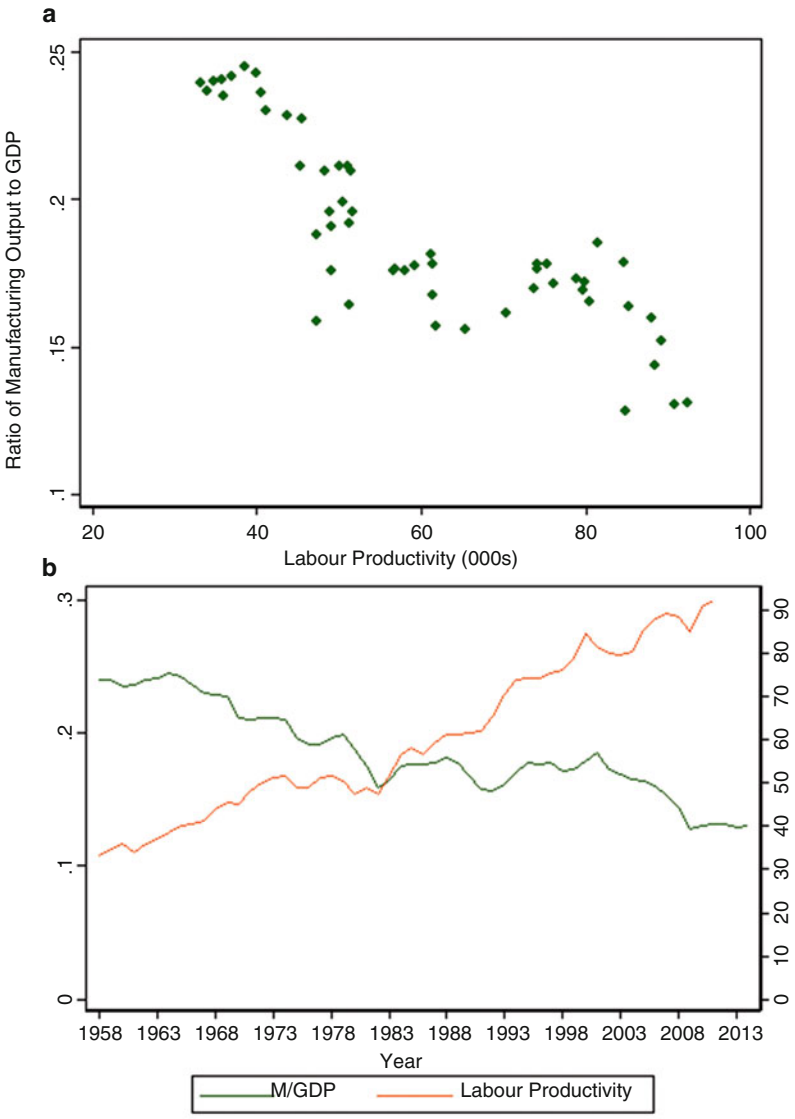


Fig. 6 (**a**) Manufacturing share of GDP versus Manufacturing labour productivity (000) 1958–2011 (**b**) Manufacturing share of GDP (1958–2014) and labour productivity (000 s – 2002 Dollars) (1958–2011) (**c**) Manufacturing share of GDP growth Versus Manufacturing labour productivity growth 1958–2011 (**d**) Manufacturing share of GDP growth (1958–2014) and Labour productivity growth (1958–2011)

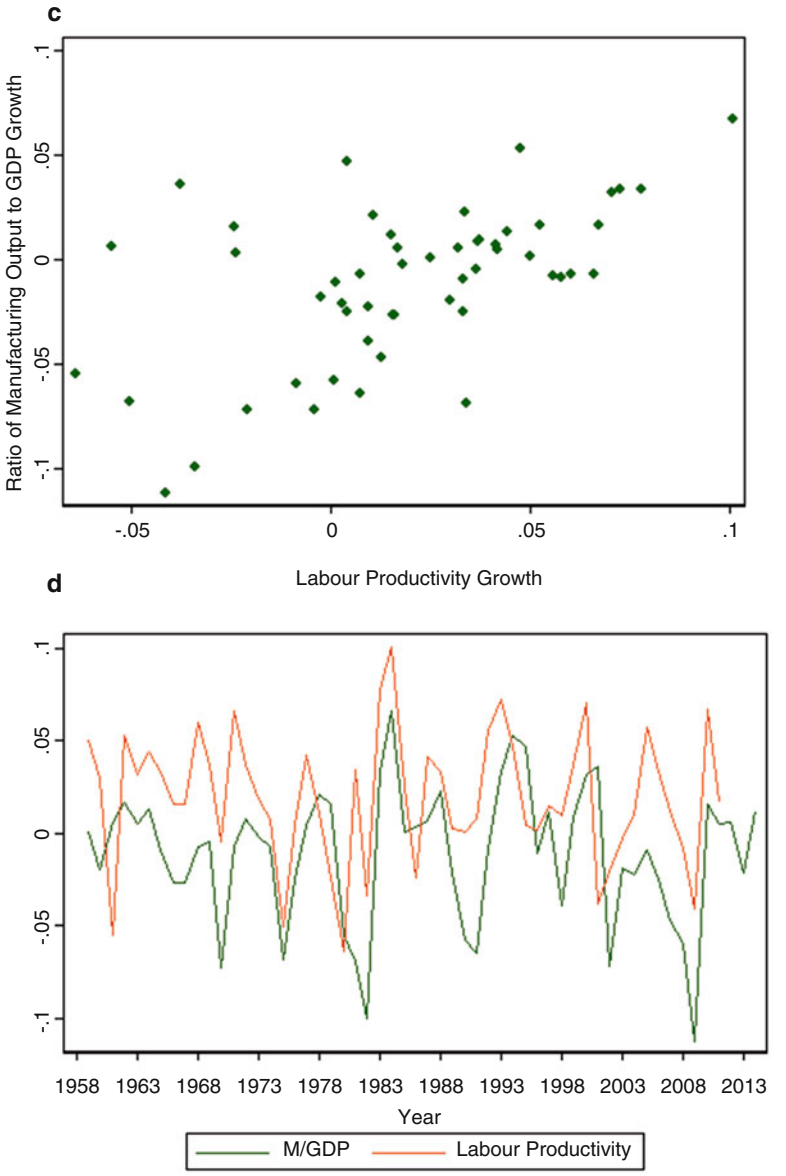


Fig. 6 (continued)

Table 2 Regressions comparing manufacturing sector's share of GDP and Labour productivity (1970–2011)

Dependent variable: Growth of manufacturing share of GDP	
Growth Labour Productivity	0.661^{***}(0.141)
Constant	-0.023^{***}(0.006)

Note: ***, **, and * indicate significance at the 1% level, 5% level and 10% level, respectively. Standard errors are in parentheses

4 The Exchange Rate Impact

Canada's manufacturing sector has declined over the course of the twentieth century in a manner similar to other developed countries though its manufacturing share of economic activity was remarkably stable between 1970 and 2000. While the recent steep decline since 2000 has been ascribed to the effect of a resource commodity boom and an appreciating Canadian dollar, there have been other periods of currency appreciation prior to 2000 which do not appear to have resulted in a shrinking manufacturing sector. Moreover, the depreciation of the Canadian dollar since 2013 does not appear to have sparked a rebound in Canadian manufacturing.

This section compares the manufacturing sector's share of GDP with exchange rate movements. The bilateral Canada-US exchange rate (CAD per USD) is used. A trade-weighted exchange rate provides an alternative to capture the movements of the Canadian dollar. Traditionally, imports from the US represent over 65% of total imports, while exports to the US represent over 75% of total exports.[7] Thus, most of the movements in the trade-weighted exchange rate are due to movements in Canada-US exchange rate.[8]

As for the effect of resources and currency appreciation on the manufacturing sector, Fig. 7a presents a plot of the relationship between the Canada-US exchange rate (CAD per USD) and the manufacturing share of GDP for the period 1926 to 2014. The diagram suggests that over the course of most of the twentieth century and into the twenty-first century, there has not been a strong relationship between a lower Canadian dollar and a larger manufacturing sector. Between parity and 1.15 CAD/USD, there is a wide variation in manufacturing to GDP ratios. Between 1.15 and 1.60 CAD/USD, there is much less fluctuation in the size of the manufacturing sector. Indeed, the manufacturing sector's share of GDP appears to be quite inelastic with respect to a Canadian dollar that depreciates to more than 1.15 CAD/USD (that is a value below about 85 cents US).

Figure 7b, c examine the relationship between Canada-US exchange rate (CAD per USD) and the manufacturing share of GDP in terms of growth rates. These figures indicate no discernible pattern or correlation between the two variables.

[7] Source: Cansim Table 228–0069.

[8] The correlation between the two variables is 0.97.

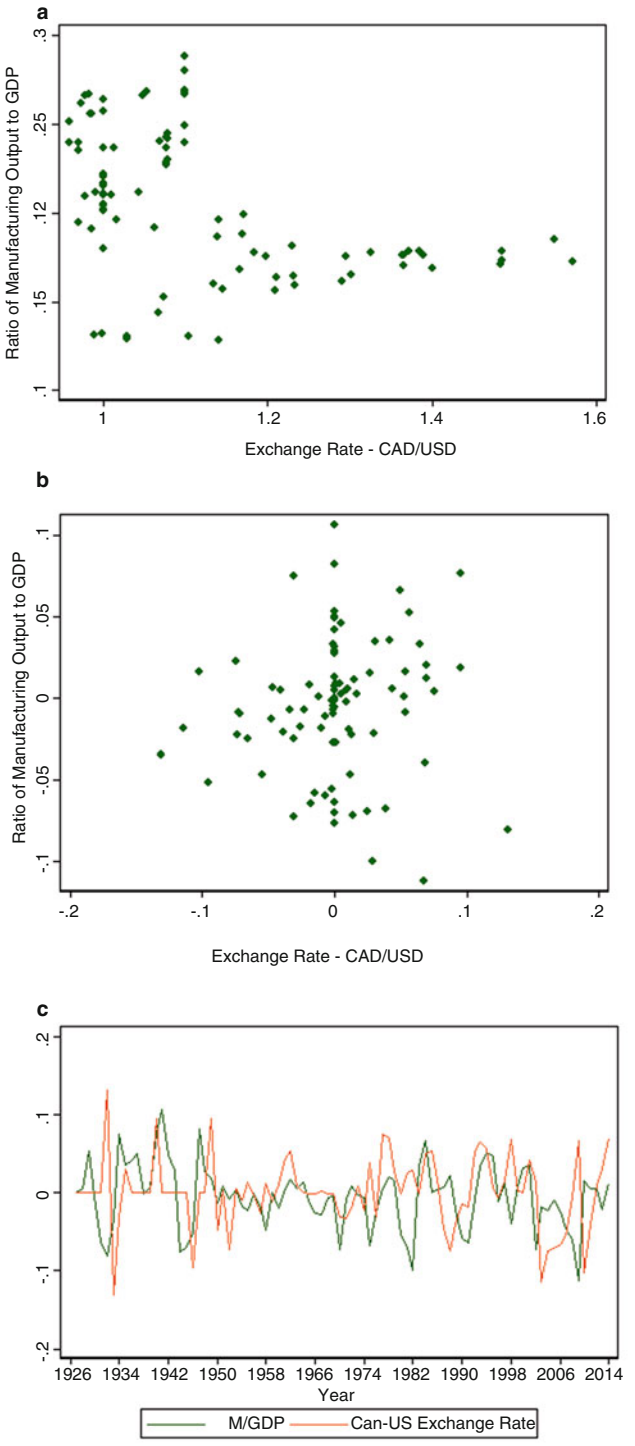


Fig. 7 (**a**) Manufacturing share of GDP versus exchange rate (CAD/USD), 1926–2014 (**b**) Growth of manufacturing share of GDP versus growth of exchange rate (CAD/USD), 1926–2014 (**c**) Growth of manufacturing share of GDP and growth of exchange rate (CAD/USD), 1926–2014

5 Conditional Analysis

To sum up the analysis thus far, the reasons for Canada's manufacturing decline are indeed complex. First, there is the "natural" evolution away from primary and secondary industry and into services which has marked the development of most advanced industrial countries. In this regard, the world's share of manufacturing output has stayed relatively constant since 1970 but the regions and countries accounting for that output have changed. Canada actually was close to the world average in terms of its manufacturing share of output for most of the latter half of the twentieth century but it has only been since 2000 that the share has significantly declined and diverged from the world average.

Second, Canada's manufacturing productivity performance has been weak relative to other countries and this may also be a factor in its relative manufacturing decline. While output per employee has increased over time, since 2000 the growth rate has slowed substantially especially compared to the United States. The decline relative to the United States is significant given the cross-border integrated nature of manufacturing – especially in the automobile sector. While the tendency has been to blame the appreciation of the Canadian dollar, the fact remains that there is not a visibly strong relationship between the manufacturing share of GDP and the value of the Canadian dollar.

Thirdly, productivity may indeed be the key variable behind the weaker performance of Canadian manufacturing especially since 2000. However, even manufacturing productivity is quite variable within Canada and may be a factor behind why Ontario and Quebec have been particularly hard hit by the recent loss of manufacturing employment. Our international productivity in manufacturing is ultimately founded on productivity across Canada's regions and the preliminary evidence suggests that manufacturing output per employee in Canada is actually high in the resource intensive provinces of Canada and particularly in western Canada which is more resource intensive than the remainder of the country. This would suggest that being resource intensive need not necessarily harm manufacturing productivity and need not be a factor in the current malaise.

Tables 3 and 4 present regression analysis for the period 1971 to 2011. Table 3 presents estimates from the regression of the level value of manufacturing share of GDP against level values of our three variables of interest: (i) G-7 (excluding Canada) share of manufacturing value added; (ii) Manufacturing labour productivity; and (iii) CAD/USD exchange rate. Table 4 provides estimates for the regressions in terms of growth rates. Each table contains seven regression specifications, which reflect the various possible combinations of the right hand side variables.

Table 3 suggests a stable relationship in level values between the manufacturing share of GDP in Canada with the three variables of interest. The coefficient on G-7 share of manufacturing value added is close to a value of two and statistically

Table 3 Regressions comparing Manufacturing sector's share of GDP (1971–2011)

	Dependent variable: Manufacturing share of GDP						
Variable:	1	2	3	4	5	6	7
G-7 (excluding Canada) Manu Share	.22(0.18)			.32(0.24)		.69***(0.22)	.98***(0.28)
Labour Productivity		−.00(0.00)		.00(0.00)	−.00(0.00)		.00(0.00)
Can-US Exchange Rate			.01(0.02)		.01(0.01)	.03***(0.01)	.03***(0.01)
Lagged Manu Share	.88***(0.08)	.94***(0.08)	.97***(0.05)	.89***(0.09)	.93***(0.08)	.75***(0.09)	.77***(0.08)
Constant	−.02(0.02)	.01(0.02)	−.01***(0.01)	−.18***(0.05)	.00(0.02)	−.18***(0.05)	−.18***(0.05)

Note: ***, **, and * indicate significance at the 1% level, 5% level and 10% level, respectively. Standard errors are in parentheses

Table 4 Regressions comparing manufacturing sector's share of GDP (1971–2011)

	Dependent variable: Growth of manufacturing Share of GDP						
Variable:	1	2	3	4	5	6	7
G-7 (excluding Canada) Manu Share Growth	.76*** (0.22)			.44**(0.21)		1.07***(0.24)	.72***(0.24)
Labour Productivity Growth		.69***(0.14)		.55***(0.15)	.69***(0.15)		.49***(0.15)
Can-US Exchange Rate Growth			.03(0.13)		.07(0.11)	.33**(0.13)	.26**(0.12)
Lagged Manu Share	−.08 (0.29)	.19 (0.27)	−.16 (0.33)	.16 (0.26)	.17 (0.28)	−.18 (0.27)	.05 (0.26)
Constant	.01(0.05)	−.06(0.05)	.02(0.06)	−.05(0.04)	−.05(0.05)	.02(0.05)	−.31***(0.07)

Note: ***, **, and * indicate significance at the 1% level, 5% level and 10% level, respectively. Standard errors are in parentheses

significant at the 1% level across the specifications.[9] Column 7 of Table 3 presents the full specification with all three independent variables. The coefficient on labour productivity is essentially zero and statistically insignificant, while the coefficient on CAD/USD exchange rate is 0.03 and statistically significant at the 1% level. Overall, these results indicate that the main drivers of the long-term movements in the Canadian manufacturing share relate to long-term movements in the manufacturing share across all the G-7 countries. This fall in manufacturing share across the G-7 is likely related to the increased importance of the service sector in these economies.

For growth rates, the relationships are less stable across various specifications as demonstrated in Table 4. Column 7 presents the full specification containing all of the variables. In this specification, the coefficients on all the variables are positive and statistically significant at the 5% level. The growth rates of the G-7 manufacturing share of value added, Canadian manufacturing labour productivity, and the CAD/USD exchange rate all provide a contribution to the growth rate in the Canadian manufacturing share of GDP. The coefficient values suggest the following. First, short-term movements of the Canadian manufacturing share follow short-term movements in the G-7 manufacturing share, but the other variables are also relevant. Second, growth in the manufacturing labour productivity leads to expansion of the manufacturing share of the Canadian economy. Finally, the positive coefficient on the CAD/USD exchange rate growth rate indicates that the appreciation of the Canadian dollar vis-à-vis the US dollar associates with a fall in the manufacturing share.

6 Manufacturing Share Across the Provinces

An examination of manufacturing and productivity over time at the provincial level is illustrative in understanding the forces behind the decline of Canadian manufacturing in light of the basic reasons advanced in this analysis. Figure 8 presents a long-term series of manufacturing as a share of GDP for Canada constructed from Statistics Canada data[10] overall as well as each of the provinces from 1961 to 2014.

The overall decline in manufacturing's share of GDP can certainly be interpreted as part of a natural evolution away from goods to service producing industries, as is the case in most developed economies. However, this natural evolution is not consistent across all the provinces. Ontario, Quebec and British Columbia have

[9]The estimates in column 1 of Table 3 do not match the estimates from Table 1, since the dependent variable is from different sources. In Table 1, the dependent variable is Canadian manufacturing share of value added taken from UN data source. In Table 3, the dependent variable is Canadian manufacturing share of GDP taken from CANSIM tables 379–0023 and 379–0028.

[10]Provincial GDP is taken from CANSIM tables 384–0015 and 384–0038. Provincial manufacturing GDP comes from CANSIM tables 379–0028, 379–0009, 379–0025 and the Historical Statistics of Canada.

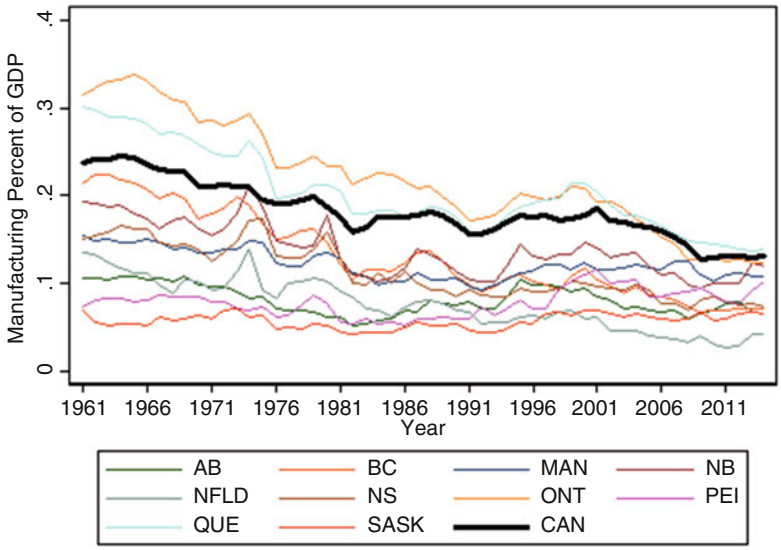


Fig. 8 Manufacturing share of GDP, Canada and provinces, 1961–2014

seen a steady decline with Ontario in particular dropping from over 30% to just over 10% over the span of 50 years. While the Prairies have historically been less manufacturing intensive than central Canada, at the same time, the declines for Saskatchewan, Alberta and Manitoba have been less pronounced. Even the Atlantic region has seen relatively stable manufacturing shares particularly since the late 1980s.

The case for either a resource curse effect or productivity malaise in explaining the performance of the manufacturing share of GDP is also mixed. Alberta, Saskatchewan and Manitoba are all much more resource intensive than either Ontario or Quebec and yet appear to have done a better job in maintaining their manufacturing sector share of GDP. If a resource curse is operating via exchange rate effects, surely it would also serve to reduce manufacturing in resource intensive provinces.

7 Conclusion

Deindustrialization and manufacturing decline have become perpetual laments in the developed world and Canada is no exception to this narrative. Recent years have seen the manufacturing heartland of Ontario and Quebec particularly hard hit by the shedding of manufacturing employment and as is inevitably the case the search for explanations has intensified. Much of the blame in recent years has focused on the

resource boom driven by Canada's energy producing provinces and the attendant effects on the appreciation of the Canadian dollar.

Our evidence suggests that manufacturing decline in Canada is indeed a complex phenomenon. To start, while it has intensified since 2000, it has generally paralleled the trend in other developed countries that over time sees an evolution away from goods production to more service intensive economies. In this sense, one could term manufacturing decline a natural evolution.

As well, Canada has seen weaker productivity growth especially since 2000 and this correlates well with the intensification of the decline during the same period. The productivity story is further reinforced by the fact that resource intensive provinces appear to have done a better job in maintaining their manufacturing sector GDP shares.

As for exchange rate effects, as an explanation this does not appear to be as well supported by our evidence suggesting the other two explanations are better. Rather than the simple story of a resource curse, Canada's manufacturing decline is more rooted in long term economic factors such as productivity growth and the evolution of economies towards service production.

References

Baldwin, R. E. (1956). Patterns of development in newly settled regions. *Manchester School of Economic and Social Studies, XXIV*, 161–179.

Baldwin, J.R., & Macdonald R. (2009). The Canadian manufacturing sector: Adapting to challenges. Statistics Canada. 11F00227M-No. 057.

Baldwin, J.R., & MacDonald, R. (2012). *Natural resources, the terms of trade, and real income growth in Canada: 1870 to 2010*. Research Paper Series, No. 11F0027M – No. 7 (Statistics Canada, Economic Analysis (EA)).

Callender, G. S. (1902). The early transportation and banking enterprises of the states in relation to the growth of corporations. *Quarterly Journal of Economics, XVII*, 111–162.

Callender, G. S. (1965[1909]). *Selections from the economic history of the United States, 1765–1860*. New York: Augustus M. Kelley.

Capeluck, E. (2015a). Explanation of the decline in manufacturing employment in Canada. Centre for the study of living standards Research Report, October, 2015–17.

Capeluck, E. (2015b). The evolution of manufacturing employment in Canada: The role of outsourcing. Centre for the study of living standards research report, October, 2015–18.

Caves, R. E., et al. (1966). Vent for surplus' models of trade and growth. In Baldwin (Ed.), *Trade, growth and the balance of payments: Essays in honor of Gottfried Haberler* (pp. 95–115). Chicago: Rand McNally.

Caves, R. E., et al. (1971). Export-led growth and the new economic history. In Bhagwati (Ed.), *Trade, balance of payments and growth* (pp. 403–442). Amsterdam: North Holland.

Corden, W. M. (1983). The economic effects of a booming sector. *International Social Science Journal, 35*, 441–454.

Corden, W. M., & Neary, J. P. (1982). Booming sector and deindustrialization in a small open economy. *Economic Journal, 92*, 825–848.

Cross, P. (2013). Dutch disease, Canadian cure. MacDonald Laurier Institute.

Cross, P. (2015). Unearthing the full economic impact of Canada's natural resources. MacDonald-Laurier Institute.

Huynh, K. P., & Petrunia, R. J. (2016). Post-entry struggle for life and pre-exit shadow of death from a financial perspective. *International Journal of the Economics of Business, 23*, 1–18.

Huynh, K. P., Jacho-Chávez, D. T., Petrunia, R. J., & Voia, M. (2011). Functional principal component analysis of density families with categorical and continuous data on Canadian entrant manufacturing firms. *Journal of the American Statistical Association, 106*, 858–878.

Innis, H. A. (1969[1956]). In M. Q. Innis (Ed.), *Essays in Canadian economic history*. Toronto: University of Toronto Press.

Innis, H. A. (1978[1940]). *The cod fisheries: The history of an international economy*. Toronto: University of Toronto Press.

Innis, H. A. (1984[1930]). *The fur trade in Canada: An introduction to Canadian economic history*. Toronto: University of Toronto Press.

Keay, I. (2007). The engine or the caboose? Resource industries and twentieth-century Canadian economic performance. *The Journal of Economic History, Vol., 67*(1), 1–32.

Sachs, J. D., & Warner, A. M. (1995) Natural resource abundance and economic growth. *NBER Working paper* 5398, p. 54

Sachs, J. D., & Warner, A. M. (1999). The big Push, natural resource booms and growth. *Journal of Development Economics, 59*, 43–76.

Sachs, J. D., & Warner, A. M. (2001). Natural resources and economic development: The curse of natural resources. *European Economic Review, 45*, 827–838.

Watkins, M. H. (1963). A staple theory of economic growth. *Canadian Journal of Economics and Political Science, XXIXE*, 141–158.

Flexible Functional Forms and Curvature Conditions: Parametric Productivity Estimation in Canadian and U.S. Manufacturing Industries

Jakir Hussain and Jean-Thomas Bernard

Abstract It is well-known that econometric productivity estimation using flexible functional forms often encounters violations of curvature conditions. However, the productivity literature does not provide any guidance on the selection of appropriate functional forms once they satisfy the theoretical regularity conditions. In this paper, we provide an empirical evidence that imposing local curvature conditions on the flexible functional forms affect total factor productivity (TFP) estimates in addition to the elasticity estimates. Moreover, we use this as a criterion for evaluating the performances of three widely used locally flexible cost functional forms—the translog (TL), the Generalized Leontief (GL), and the Normalized Quadratic (NQ)—in providing TFP estimates. Results suggest that the NQ model performs better than the other two functional forms in providing TFP estimates.

Keywords Technical change · Productivity · Flexible functional forms · Translog (TL) · Generalized Leontief (GL) · Normalized Quadratic (NQ) cost function · Concavity

JEL classification: C22, F33

We gratefully acknowledge the financial assistance provided by the Social Sciences and Humanities Research Council of Canada (SSHRC). We thank Yazid Dissou for sharing the Canadian KLEMS data set obtained from Statistics Canada. We would like to thank Samuel Gamtessa, Lynda Khalaf, Yazid Dissou, Pierre Brochu and an anonymous referee for valuable comments and suggestions. We would also like to thank the participants at North American Productivity Workshop IX, June 2016, Quebec City, and 49th Annual Conference of Canadian Economics Association, May 2015, Toronto, for helpful comments and discussions.

J. Hussain (✉)
Department of Economics, University of Ottawa, Ottawa, ON, Canada, K1N 6N5
e-mail: jhussain@uOttawa.ca

J.-T. Bernard
Department of Economics, University of Ottawa, Ottawa, ON, Canada
e-mail: jbernar3@uOttawa.ca

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_10

1 Introduction

The standard econometric approach to modelling technical change, introduced by Binswanger (1974a,b), involves representing the rate and biases through constant time trends in a flexible functional form and estimating the unknown parameters using econometric methods.[1] See Jin and Jorgenson (2010) for a list of studies that rely on this widely used approach for modelling technical change. It is well known that among the regularity conditions—positivity, monotonicity, and curvature—that are implied by economic theory, curvature conditions are often violated in empirical applications of flexible functional forms (see Diewert and Wales 1987; Ryan and Wales 2000).

Theoretical curvature properties are crucial, especially, in estimating the flexible functional forms. For example, Diewert and Wales (1987) noted that it is necessary for the estimated production and utility functions used in applied general equilibrium models to globally satisfy the theoretical curvature conditions. Empirical rejection of concavity by the estimated cost function casts a doubt on the underlying true model of production as it could lead to a non-continuous input demand function. Moreover, any inferences based on this result would be unconvincing since the input demand functions derived from the cost function may not be cost minimizing due to the violation of curvature property. This casts a serious doubt on the assumption that firms in the sample are cost minimizers.

No parametric restrictions can ensure global curvature conditions while maintaining flexibility in the translog (TL) and the Generalized Leontief (GL) functional forms as this property is data dependent. Violation of curvature conditions by the TL and the GL has led Diewert and Wales (1987) to develop a more complex locally flexible functional form—the Normalized Quadratic (NQ) (see Diewert and Wales 1987)—which allows imposing global curvature conditions and maintaining flexibility at the same time. However, instead of imposing global concavity on the TL and the GL cost functions, Ryan and Wales (2000) propose a method to impose it locally, at a chosen reference point. They show that their procedure of curvature imposition does not destroy the flexibility property of the functional forms and it leads to satisfaction of curvature property at all data points for their data set. Since they do not find any impact on productivity estimates they conclude that the effect of imposing concavity is limited, in their case, only to the price responses. However, the productivity literature does not provide any guidance on the selection of appropriate flexible functional forms once they satisfy all theoretical regularity conditions (see Feng and Serletis 2008).

[1] Slade (1989) criticizes the traditional method of modelling the state of technology by including time trend in the production or cost function and, instead, suggests the use of state-space approach through the Kalman filter in estimating technical change. More recently, Jin and Jorgenson (2010) replaces the constant time trend by latent variables and use the Kalman filter to estimate the latent variables in the translog (TL) model.

In this paper we provide empirical evidence that imposing local concavity on the TL and the GL cost functions affects the productivity estimates, in addition to its effect on the elasticity estimates which has attracted the attention in the literature thus far. In doing so, we employ three well-known locally flexible functional forms—the TL, the GL, and the NQ cost functions, and present an empirical comparison and evaluation of the effectiveness of these cost functions in providing total factor productivity (TFP) estimates when they satisfy all theoretical regularity conditions.[2] We use the difference in TFP estimates, with and without curvature constraint, as a criterion for comparing performances of different functional forms. In estimating the models we utilize manufacturing KLEM (capital, labour, energy, and material) data covering the period from 1961 to 2003 for Canada and the U.S. Moreover, we follow Ryan and Wales (2000) to impose local curvature conditions on the TL and the GL models, and for the NQ we impose global curvature conditions following Diewert and Wales (1987).

While comparing the functional forms in providing TFP estimates, we provide examples of all three possible scenarios: First, the cases where all three regularity conditions are satisfied without curvature being imposed; second, the cases where curvature conditions are not satisfied even with local curvature being imposed; finally, the cases where imposing curvature resulted in the satisfaction of regularity conditions. We also provide the price elasticity estimates for factors of production in the U.S. and Canadian manufacturing industries.

Feng and Serletis (2008) estimate TFP in the U.S. manufacturing industry using four flexible functional forms. In addition to the three functional forms that we use in this study, they also estimate the asymptotically ideal model (AIM) cost function. Although they impose concavity on the three locally flexible functional forms using the same techniques that we use in this study, only the NQ model satisfies curvature conditions. As a result, they provide a comparison between the NQ and the AIM cost functions—the only models that satisfy all regularity conditions. Using a smoothed Fisher TFP index as the benchmark they conclude that the AIM, with curvature imposed, performs better in estimating TFP. However, there is not a single case where concavity is satisfied without curvature being imposed and as a result they are unable to analyze this aspect. Moreover, since they fail to include the TL and the GL functional forms in the analysis they could not analyze the impact of local curvature imposition. Based on a different criterion we provide comparisons between all three locally flexible functional forms that we consider in this study. We also compare the effects of local and global curvature imposition while Feng and Serletis (2008) evaluate the effect of global curvature imposition. Our estimation results provide us the opportunity to analyze all three possible scenarios mentioned earlier, which is not the case in Feng and Serletis (2008).

[2]Fisher et al. (2001) provide an empirical evaluation of the performances of eight flexible functional forms in the context of consumer demand.

Our empirical findings support the result provided by Feng and Serletis (2008) that imposing local curvature conditions on the TL and the GL does not always assure theoretical regularity conditions at all data points in the sample. Moreover, using the estimation results we show that when curvature condition is met by imposition, the NQ model performs better in providing TFP estimates than the TL and GL models, at least for our data sets. However, the GL model performs equally well when all regularity conditions are satisfied without curvature being imposed. Our findings also provide evidence that local curvature imposition on the TL and the GL models affect the productivity estimates. Based on our results we argue that since concavity is not guaranteed in the TL and the GL models even with curvature being imposed, functional forms with global curvature conditions appear to be a better choice for econometric productivity estimation.

The rest of the paper is organized as follows. Section 2 reviews the theoretical background on different approaches to productivity measurement. Section 3 presents the functional forms and relevant techniques for imposing curvature conditions while Sect. 4 discusses the index number techniques used in this study. Section 5 describes the data sets and outlines the empirical estimation strategies. Section 6 reports the estimation results, and finally, Sect. 7 concludes.

2 Theoretical Background

2.1 Productivity Measurement

Recent shift in attention to the rate and biases of technical change has put the econometric approach to productivity measurement in the forefront of empirical productivity analysis even though the most commonly utilized approach is the index number technique.[3] See, for example, Acemoglu (2002, 2007), Jaffe et al. (2003), Jin and Jorgenson (2010) for different applications of the biases of technical change. As opposed to the index number approach, it is ideal to have a productivity measure that would also shed some light on the production structure, for example, the factor biases and the elasticities of factor substitution. Accurate information on these estimates is vital for policy making, in particular energy policies. See, for example, León-Ledesma et al. (2010) for a discussion on the importance of parameters of the elasticity of substitution and the direction of technical change. In what follows, we briefly discuss the index number and the econometric approaches to productivity measurement.[4]

[3] See Hulten (2001) for a discussion on the historical development of quantitative analysis of productivity. For a brief discussion on various approaches to productivity measurement see Feng and Serletis (2008). Using simulation Van Biesebroeck (2007) provides a discussion on the robustness of productivity estimates obtained by different measurement approaches.

[4] As in Feng and Serletis (2008), this section builds heavily on standard notations in the literature, mainly from Berndt (1991).

With the index number approach, the difference between the rate of change of output and a share weighted index of rates of change in inputs provides a measure of the rate of technical change,

$$\frac{\partial y/\partial t}{y} - \sum_{i=1}^{n} s_i \frac{\partial x_i/\partial t}{x_i} \tag{1}$$

where inputs $\mathbf{x} \geq 0$ are used to produce output y, and $s_i \equiv \frac{p_i x_i}{C}$ is the share of input x_i in total cost C.

The econometric approach typically involves estimating a production or cost function which is then differentiated with respect to time. Technical change is associated with any temporal shift in the production or cost frontier, and under constant returns to scale both production and cost function approaches would yield equivalent results. With the cost function approach it is assumed that there exists a cost function,

$$C = C(\mathbf{p}, y, t), \tag{2}$$

that relates the vector of factor prices $\mathbf{p}$, output y, and a proxy of technical change t with total cost C.[5] The cost function in (2) is a solution to the following cost minimization problem,

$$\min_{\mathbf{x}} \mathbf{px} \quad s.t. \quad y \leq f(\mathbf{x}, t),\ \mathbf{x} \geq \mathbf{0},$$

and is the corresponding dual of the following strictly monotonic, strictly quasi-concave, and continuously twice differentiable production function,

$$y = f(\mathbf{x}, t). \tag{3}$$

To be able to successfully represent the underlying production process in (3), the cost function (2) must be nonnegative, non-decreasing in y and $\mathbf{p}$, and linear homogeneous and concave in $\mathbf{p}$. Under constant returns to scale, Eq. (2) becomes

$$C(\mathbf{p}, y, t) = yc(\mathbf{p}, t) \tag{4}$$

where c is the corresponding unit cost function. Invoking Shephard's Lemma yields the optimal factor demand equations,

$$x_i = \partial C(\mathbf{p}, y, t)/\partial p_i, \quad i = 1, \ldots, n. \tag{5}$$

[5]For notational simplicity we suppress the time subscripts.

Rate of technical change in the cost function approach is measured in the following way,

$$TFP = -\frac{\partial \ln C(\mathbf{p}, y, t)/\partial t}{\partial \ln C(\mathbf{p}, y, t)/\partial \ln y} \tag{6}$$

where $\partial \ln C(\mathbf{p}, y, t)/\partial t$ represents the rate of cost reduction and $\partial \ln C(\mathbf{p}, y, t)/\partial \ln y$ represents the inverse of rate of returns to scale which equals to unity under constant returns to scale,

$$TFP = -\frac{\partial \ln C(\mathbf{p}, y, t)}{\partial t} = -\frac{\partial C(\mathbf{p}, y, t)/\partial t}{C}. \tag{7}$$

Hence, an upward shift in the production function is equal to an equivalent downward shift in the cost function under constant returns to scale. We can easily obtain the TFP estimates by estimating the parameters of C once we assume the specific functional form for C in (2).

The effects of technological change on factor use, which is often referred to as the input bias due to technical change, can also be measured as follows

$$\tau_i = \frac{\partial \ln x_i(\mathbf{p}, y, t)}{\partial t}. \tag{8}$$

They provide us with the information about how the usage of inputs changes as a result of technical change. With regards to the direction of bias, if $\tau_i > 0 (< 0)$, then technical change is factor i using (saving). Moreover, the input price elasticities are computed as

$$\eta_{ij} = \frac{\partial \ln x_i(\mathbf{p}, y, t)}{\partial \ln p_j} = \frac{\partial x_i(\mathbf{p}, y, t)}{\partial p_j} \frac{p_j}{x_i(\mathbf{p}, y, t)}. \tag{9}$$

2.2 Curvature Conditions

A crucial requirement about the cost function is that it must be concave in prices. Necessary and sufficient condition for concavity is that the Hessian matrix of the cost function be negative semi-definite. To check for concavity, eigenvalues of the estimated Hessian matrix of the cost functions are computed at each point in the sample space. Morey (1986) provides an excellent overview of the curvature conditions and checking process for empirical applications of different flexible functional forms.

As mentioned earlier, curvature properties are often violated in empirical applications of flexible functional forms. One way to deal with these violations is to perform some simple checking procedures. For example, Caves et al. (1980) check whether the parameter generating positive own price elasticity is significantly different

from a value of the parameter generating non-positive elasticity estimate. Another simple procedure is adopted by Slade (1989) which involves checking whether the concavity violations occurred by chance. This procedure uses a technique developed by Geweke (1986) and involves Monte Carlo simulation method to generate replications of the estimated coefficients from a multivariate normal distribution.

The other available option is to impose local or global curvature restrictions. Curvature restrictions are imposed in many different contexts. For applications, in consumer theory, see, for example, Ryan and Wales (1998), Moschini (1999), in case of production theory, see Gallant and Golub (1984), Terrell (1996), and Ryan and Wales (2000), and in the context of GNP functions, see Kohli (1992). Global curvature restrictions can be imposed using Cholesky decomposition on the NQ cost function without destroying its flexibility property. For a discussion on the restrictions and implementation technique, see Diewert and Wales (1987) and Wiley et al. (1973). However, imposing global restrictions on the TL and the GL cost functions destroys their flexibility property.

Instead of imposing global concavity, Ryan and Wales (2000) propose a method to impose it locally at a chosen reference point. Using the same dataset as Diewert and Wales (1987), they show that their procedure of imposing local curvature guarantees concavity at the data point where it is imposed without destroying the flexibility of the functional form. Although imposing concavity on a single observation does not guarantee concavity for other data points, they hope that a judicious choice of point of imposition may lead to satisfaction of concavity at most or all data points. This procedure of local curvature imposition provides the expected concavity coverage for their data set; however, this result is not universal. Other techniques for imposing local curvature conditions include the general computational methods used by Lau (1978) and Gallant and Golub (1984), and the Bayesian approach used by Chalfant and Wallace (1992), Terrell (1996), and Griffiths et al. (2000).

In what follows, we take the econometric approach to productivity measurement for three Canadian and two U.S. energy intensive industries and provide a comparison between three widely used locally flexible cost functional forms—the TL, the GL, and the NQ cost functions. For all cost functional forms, we assume cross-equation symmetry restrictions, linear homogeneity in prices as well as constant returns to scale in the production process as a maintained hypothesis. Obviously one can test for whether these assumptions actually hold.

3 Locally Flexible Functional Forms

3.1 The TL Cost Function

If we assume that the unit cost function in (4) takes the TL functional form (Christensen et al. 1971, 1973), then we get

$$\ln C(\mathbf{p}, y, t) = \beta_0 + \ln y + \sum_{i=1}^{n} \beta_i \ln p_i + \frac{1}{2} \sum_{i=1}^{n} \beta_{ii} [\ln p_i]^2 + \frac{1}{2} \sum_{i \neq j}^{n} \sum_{j=1}^{n} \beta_{ij} \ln p_i \ln p_j + \sum_{i=1}^{n} \beta_{it} t \ln p_i + \beta_t t + \frac{1}{2} \beta_{tt} t^2. \tag{10}$$

Together with the assumption of symmetry—that is, $\beta_{ij} = \beta_{ji}$—homogeneity of degree one in prices imposes the following constraints

$$\sum_{i=1}^{n} \beta_i = 1, \quad \sum_{i=1}^{n} \beta_{ij} = \sum_{j=1}^{n} \beta_{ij} = \sum_{i=1}^{n} \beta_{it} = 0. \tag{11}$$

The corresponding factor cost share equations are

$$s_i = \frac{p_i x_i}{C} = \beta_i + \beta_{ii} \ln p_i + \sum_{j \neq i}^{n} \beta_{ij} \ln p_j + \beta_{it} t. \tag{12}$$

Equation (12) imposes another adding-up restriction, $\sum_{i=1}^{n} s_i = 1$, which already holds through the assumption of linear homogeneity in prices. Since all parameters of the share equations are also present in the cost function, we can directly estimate (10). However, joint estimation of (10) and (12) as a system of equations reduces possible high degree of multicollinearity in the independent variables, and increases efficiency and degrees of freedom available.

However, as mentioned earlier, in empirical applications failure of the TL to satisfy the regularity conditions is very common. To overcome this problem local concavity can be imposed at a chosen reference point following the route suggested by Ryan and Wales (2000). Diewert and Wales (1987) show that the Hessian of the TL cost function will be negative semidefinite, providing $C(\mathbf{p}, y, t) > 0$, if and only if the following matrix is negative semidefinite

$$\mathbf{H} = \mathbf{B} - \mathbf{S}^n + \mathbf{S}\mathbf{S}' \tag{13}$$

where $\mathbf{B} = [\beta_{ij}]$ is the $n \times n$ symmetric matrix of β_{ij}, $\mathbf{S} = [s_1, \ldots, s_n]'$ is the vector of input shares, and $\mathbf{S}^n$ is the $n \times n$ diagonal matrix of input shares.

To impose local concavity, Eq. (10) is rewritten as

$$\ln C(\mathbf{p}, y, t) = \beta_0 + \ln y + \sum_{i=1}^{n} \beta_i \ln p_i + \frac{1}{2} \sum_{i=1}^{n} \beta_{ii} [\ln p_i]^2 + \frac{1}{2} \sum_{i \neq j}^{n} \sum_{j=1}^{n} \beta_{ij} \ln p_i \ln p_j + \sum_{i=1}^{n} \beta_{it} t \ln p_i + \beta_t [t - t^*] + \frac{1}{2} \beta_{tt} [t - t^*]^2 \tag{14}$$

where t^* is the chosen reference point—that is, the point of imposition of local concavity—where all prices are normalized to one. The corresponding input share equations are

$$s_i = \beta_i + \beta_{ii} \ln p_i + \sum_{j \neq i}^{n} \beta_{ij} \ln p_j + \beta_{it}[t - t^*]. \tag{15}$$

Normalizing all input prices to one at t^* makes $s_i = \beta_i$ for all i at this data point. The ijth element of $\mathbf{H}$ evaluated at t^* is

$$\mathbf{H}_{ij} = \beta_{ij} + \beta_i \beta_j - \beta_i \delta_{ij} \qquad i, j = 1, \ldots, n, \tag{16}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Curvature is imposed at the reference point, t^*, by setting $\mathbf{H} = -\mathbf{A}\mathbf{A}'$, where $\mathbf{A}$ is a lower triangular matrix with elements a_{ij} for $i \geq j$ and 0 elsewhere. Now solving for $\mathbf{A}$ in (16) gives us

$$\beta_{ij} = -(\mathbf{A}\mathbf{A}')_{ij} + \beta_i \delta_{ij} - \beta_i \beta_j \qquad i, j = 1, \ldots, n, \tag{17}$$

where $(\mathbf{A}\mathbf{A}')_{ij}$ is the ijth element of $\mathbf{A}\mathbf{A}'$.

Equation (17) gives us the following relationships, in the case of four factors:

$$\begin{aligned}
\beta_{11} &= -a_{11}^2 + \beta_1 - \beta_1^2, & \beta_{12} &= -a_{11}a_{21} - \beta_1\beta_2, \\
\beta_{13} &= -a_{11}a_{31} - \beta_1\beta_3, & \beta_{14} &= -a_{11}a_{41} - \beta 1\beta_4, \\
\beta_{22} &= -(a_{21}^2 + a_{22}^2) + \beta_2 - \beta_2^2, & \beta_{23} &= -(a_{21}a_{31} + a_{22}a_{32}) - \beta_2\beta_3, \\
\beta_{24} &= -(a_{21}a_{41} + a_{22}a_{42}) - \beta_2\beta 4, & \beta_{33} &= -(a_{31}^2 + a_{32}^2 + a_{33}^2) + \beta_3 - \beta_3^2, \\
\beta_{34} &= -(a_{31}a_{41} + a_{32}a_{42} + a_{33}a_{43}) - \beta_3\beta_4, & \beta_{44} &= -(a_{41}^2 + a_{42}^2 + a_{43}^2 + a_{44}^2) + \beta_4 - \beta_4^2.
\end{aligned}$$

Replacing the elements of $\mathbf{B} = [\beta_{ij}]$ in (14) and (15) by the above relationships and estimating a_{ij} will ensure that the TL cost function will be concave at the normalization point t^*, and may also encompass concavity at other data points in the sample. However, this replacement makes the system of equations nonlinear in parameters a_{ij}.

Equation (7) yields the rate of technical change for the TL as

$$TFP = -\frac{\partial \ln C(\mathbf{p}, y, t)}{\partial t} = -\left(\beta_t + \beta_{tt} t + \sum_{i=1}^{n} \beta_{it} \ln p_i\right). \tag{18}$$

Moreover, following Jin and Jorgenson (2010), we decompose the rate of technical change into autonomous and induced technical change components. Together, the first two parts on the right hand side, which depends only on changes in the level of technology, gives us the rate of autonomous technical change. The last part on the right hand side of (18), which depends on the prices as well as the biases of technical

change, measures the contribution to rate of productivity growth due to the biased effect of technical change and the change in relative input prices. We refer to this as the rate of induced technical change.

Own- and cross-price elasticities are calculated as

$$\eta_{ij} = \frac{\widehat{\beta}_{ij} + \widehat{s_i}\widehat{s_j}}{\widehat{s_i}}, \qquad \text{for } i \neq j, \tag{19}$$

$$\eta_{ii} = \frac{\widehat{\beta}_{ii} + \widehat{s}_i^2 - \widehat{s_i}}{\widehat{s_i}}, \qquad i, j = 1, \ldots, n. \tag{20}$$

3.2 The Generalized Leontief Cost Function

If we choose the functional form in (4) to be the GL cost function (Diewert and Wales 1987), we get

$$C(\mathbf{p}, y, t) = y \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij} (p_i p_j)^{\frac{1}{2}} + \sum_{i=1}^{n} \beta_{it} p_i t + \sum_{i=1}^{n} \gamma_{it} p_i t^2 \right) \tag{21}$$

where $\beta_{ij} = \beta_{ji}$. Using (5) we get the corresponding optimal input-output demand equations as follows,

$$a_i = \frac{x_i}{y} = \sum_{j=1}^{n} \beta_{ij} p_j^{\frac{1}{2}} p_i^{-\frac{1}{2}} + \beta_{it} t + \gamma_{it} t^2, \qquad i = 1, \ldots, n. \tag{22}$$

There is no intercept term in (21) due to the assumption of constant returns to scale and hence all the parameters in (21) can be obtained by estimating only (22). When $i = j$ in (22), β_{ii} becomes a constant term in the ith input-output equation and if $\beta_{ij} = 0$ for all $i, j, i \neq j$, (22) becomes independent of relative input prices and all the cross-price elasticities become zero.

Necessary and sufficient condition for concavity is that the Hessian matrix of (21) be negative semi-definite. The elements of the Hessian matrix for (21) are as follows

$$\begin{aligned} \mathbf{H}_{ij} &= \frac{1}{2} \beta_{ij} (p_i p_j)^{-\frac{1}{2}} && i \neq j \\ &= -\frac{1}{2} \sum_{j \neq i}^{n} \beta_{ij} (p_j / p_i)^{\frac{1}{2}} (1/p_i) && i = j. \end{aligned} \tag{23}$$

Following Ryan and Wales (2000), we reparametrize Eq. (21) to impose local concavity. First, we set $\sum_{k=1}^{n} \beta_{ik} = 0$ for all i, and then we add $(\sum_{i=1}^{n} p_i d_i)y$ to the right hand side of (21) which will introduce n new parameters. Therefore, the new set of input-output demand equations are

$$a_i = \frac{x_i}{y} = \sum_{j=1}^{n} \beta_{ij} p_j^{\frac{1}{2}} p_i^{-\frac{1}{2}} + \beta_{it} t + \gamma_{it} t^2 + d_i \qquad i = 1, \ldots, n. \tag{24}$$

Second, we normalize all prices and output to one at the reference point and, as a result, the ijth element of $\mathbf{H}$ evaluated at this data point becomes

$$\begin{aligned} \mathbf{H}_{ij} &= \frac{\beta_{ij}}{2} && i \neq j \\ &= -\frac{1}{2} \sum_{j \neq i}^{n} \beta_{ij} = \frac{\beta_{ii}}{2} && i = j. \end{aligned} \tag{25}$$

Finally, we set $\beta_{ij} = -(\mathbf{DD}')_{ij}$, where $(\mathbf{DD}')_{ij}$ is the ijth element of $\mathbf{DD}'$, and $\mathbf{D}$ is a lower triangular matrix with elements d_{ij} for $i \geq j$ and 0 elsewhere. This gives us the following relationships between β_{ij} and d_{ij}, in the case of four factors,

$$\begin{aligned} \beta_{11} &= -d_{11}^2, & \beta_{12} &= -d_{11}d_{21}, \\ \beta_{13} &= -d_{11}d_{31}, & \beta_{14} &= -d_{11}d_{41}, \\ \beta_{22} &= -(d_{21}^2 + d_{22}^2), & \beta_{23} &= -(d_{21}d_{31} + d_{22}d_{32}), \\ \beta_{24} &= -(d_{21}d_{41} + d_{22}d_{42}), & \beta_{33} &= -(d_{31}^2 + d_{32}^2 + d_{33}^2), \\ \beta_{34} &= -(d_{31}d_{41} + d_{32}d_{42} + d_{33}d_{43}), & \beta_{44} &= -(d_{41}^2 + d_{42}^2 + d_{43}^2 + d_{44}^2). \end{aligned} \tag{26}$$

Replacing the elements of $\mathbf{B} = [\beta_{ij}]$ in (24) by the relationships in (26) and estimating d_{ij} will guarantee that the estimated GL cost function will be concave at the normalization point and it may also lead to the satisfaction of concavity at other data points in the sample.

For the GL specification we compute the price elasticities as

$$\eta_{ij} = \frac{1}{2} \frac{\beta_{ij}(p_i/p_j)^{-\frac{1}{2}}}{a_i}, \qquad \text{for } i \neq j, \tag{27}$$

$$\eta_{ii} = -\frac{1}{2} \frac{\sum_{j \neq i}^{n} \beta_{ij}(p_i/p_j)^{-\frac{1}{2}}}{a_i}. \tag{28}$$

3.3 The Normalized Quadratic Cost Function

If we consider the functional form in (4) to be the NQ cost function (Diewert and Wales 1987), we get

$$C(\mathbf{p}, y, t) = y\left[\sum_{i=1}^{n}\beta_i p_i + \frac{1}{2}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij}p_i p_j}{\sum_{i=1}^{n}\theta_i p_i} + \sum_{i=1}^{n}\beta_{it}p_i t + \sum_{i=1}^{n}\gamma_{it}p_i t^2\right] \quad (29)$$

where $\beta_{ij} = \beta_{ji}$. Equation (29) can be rewritten as

$$C(\mathbf{p}, y, t) = y\left[\sum_{i=1}^{n}\beta_i p_i + g(\mathbf{p}) + \sum_{i=1}^{n}\beta_{it}p_i t + \sum_{i=1}^{n}\gamma_{it}p_i t^2\right] \quad (30)$$

where $g(\mathbf{p}) \equiv \frac{\mathbf{p}'\mathbf{B}\mathbf{p}}{2\boldsymbol{\theta}'\mathbf{p}}$. $\mathbf{B} \equiv [\beta_{ij}]$ is a $n \times n$ symmetric matrix, and $\boldsymbol{\theta} = [\theta_1, \ldots, \theta n] > \mathbf{0}$ is a vector of nonnegative constants, not all equal to zero. Usually $\boldsymbol{\theta}$ is predetermined, and we set θ_i equal to the sample average values of the respective inputs. In order to identify all parameters in the model, n extra restrictions on the elements of $\mathbf{B}$ are imposed as

$$\mathbf{B}\mathbf{p}^* = \mathbf{0}, \quad (31)$$

for some chosen $\mathbf{p}^* > \mathbf{0}$.

Application of (5) yields the following system of n equations

$$\frac{x_i}{y} = \beta_i + \sum_{j=1}^{n}\beta_{ij}\frac{p_i}{\sum_{i=1}^{n}\theta_i p_i} - \frac{1}{2}\theta_i\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij}\frac{p_i}{\sum_{i=1}^{n}\theta_i p_i}\frac{p_j}{\sum_{j=1}^{n}\theta_j p_j}\right) + \beta_{it}t + \gamma_{it}t^2. \quad (32)$$

Equation (32) can also be expressed as

$$\frac{x_i}{y} = \beta_i + \frac{\sum_{j=1}^{n}\beta_{ij}p_i}{\boldsymbol{\theta}'\mathbf{p}} - \frac{\theta_i}{2}\frac{\mathbf{p}'\mathbf{B}\mathbf{p}}{(\boldsymbol{\theta}'\mathbf{p})^2} + \beta_{it}t + \gamma_{it}t^2. \quad (33)$$

Furthermore, we assume $\mathbf{p}^* = \mathbf{1_n}$ which, in terms of (31), implies $\sum_{j=1}^{n}\beta_{ij} = 0$. With these n restrictions on matrix $\mathbf{B}$ and after denoting $w_i = \frac{p_i}{\sum_{j=1}^{n}\theta_i p_i}$, the system of factor demand equations (32), in the case of four factors, can be expressed as,

$$\frac{x_1}{y} = \beta_{11}\left[(w_1 - w_4) - \frac{\theta_1}{2}(w_1 - w_4)^2\right] + \beta_{12}\left[(w_2 - w_4) - \theta_1(w_1 - w_4)(w_2 - w_4)\right]$$
$$+ \beta_{13}\left[(w_3 - w_4) - \theta_1(w_1 - w_4)(w_3 - w_4)\right] + \beta_{22}\left[\frac{\theta_1}{2}(w_2 - w_4)^2\right]$$

$$+\beta_{23}\Big[-\theta_1(w_2-w_4)(w_3-w_4)\Big]+\beta_{33}\Big[-\frac{\theta_1}{2}(w_3-w_4)^2\Big]+\beta_1+\beta_{1t}t+\gamma_{1t}t^2, \tag{34}$$

$$\frac{x_2}{y}=\beta_{11}\Big[-\frac{\theta_2}{2}(w_1-w_4)^2\Big]+\beta_{12}\Big[(w_1-w_4)-\theta_2(w_1-w_4)(w_2-w_4)\Big]$$
$$+\beta_{13}\Big[-\theta_2(w_1-w_4)(w_3-w_4)\Big]+\beta_{22}\Big[(w_2-w_4)-\frac{\theta_2}{2}(w_2-w_4)^2\Big]$$
$$+\beta_{23}\Big[(w_3-w_4)-\theta_2(w_3-w_4)(w_2-w_4)\Big]+\beta_{33}\Big[\frac{\theta_2}{2}(w_3-w_4)^2\Big]$$
$$+\beta_2+\beta_{2t}t+\gamma_{2t}t^2, \tag{35}$$

$$\frac{x_3}{y}=\beta_{11}\Big[-\frac{\theta_3}{2}(w_1-w_4)^2\Big]+\beta_{12}\Big[-\theta_3(w_1-w_4)(w_2-w_4)\Big]$$
$$+\beta_{13}\Big[(w_1-w_4)-\theta_3(w_1-w_4)(w_3-w_4)\Big]+\beta_{22}\Big[\frac{\theta_3}{2}(w_2-w_4)^2\Big]$$
$$+\beta_{23}\Big[(w_2-w_4)-\theta_3(w_2-w_4)(w_3-w_4)\Big]+\beta_{33}\Big[(w_3-w_4)-\frac{\theta_3}{2}(w_3-w_4)^2\Big]$$
$$+\beta_3+\beta_{3t}t+\gamma_{3t}t^2, \tag{36}$$

$$\frac{x_4}{y}=\beta_{11}\Big[-(w_1-w_4)-\frac{\theta_4}{2}(w_1-w_4)^2\Big]+\beta_{12}\Big[-(w_2-w_4)-\theta_4(w_1-w_4)(w_2-w_4)\Big]$$
$$+\beta_{13}\Big[-(w_3-w_4)-\theta_4(w_1-w_4)(w_3-w_4)\Big]+\beta_{22}\Big[(w_2-w_4)-\frac{\theta_4}{2}(w_2-w_4)^2\Big]$$
$$+\beta_{23}\Big[(w_2-w_4)-(w_3-w_4)-\theta_4(w_1-w_4)(w_3-w_4)\Big]$$
$$+\beta_{33}\Big[(w_3-w_4)-\frac{\theta_4}{2}(w_2-w_4)^2\Big]+\beta_4+\beta_{4t}t+\gamma_{4t}t^2. \tag{37}$$

Estimates of β_i, β_{it}, and the elements of matrix $\mathbf{B}$, except for the parameters β_{i4} ($i=1,2,3,4$), are obtained by estimating the system of input-output equations (34), (35), (36), and (37). β_{i4} can then be recovered from the imposed restrictions.

Global concavity for the NQ cost function requires that the estimated $\mathbf{B}$ matrix is negative semidefinite, provided that $\boldsymbol{\theta} > \mathbf{0}$. However, in empirical applications the estimated $\mathbf{B}$ matrix may not be negative semidefinite and if this turns out to be the case, Diewert and Wales (1987) show that global concavity on the NQ function can be imposed without destroying its flexibility by using the technique suggested by Wiley et al. (1973).

To impose global concavity we set

$$\mathbf{B}=-\mathbf{A}\mathbf{A}',$$

where $\mathbf{A}$ is a lower triangular matrix, with elements a_{ij} for $i \geq j$ and 0 elsewhere, that satisfies

$$\mathbf{A}'\mathbf{p}^*=\mathbf{0}_n.$$

This gives us the following relationships between β_{ij} and a_{ij},

$$\begin{aligned} \beta_{11} &= -a_{11}^2, & \beta_{12} &= -a_{11}a_{21}, \\ \beta_{13} &= -a_{11}a_{31}, & \beta_{22} &= -(a_{21}^2 + a_{22}^2), \\ \beta_{23} &= -(a_{21}a_{31} + a_{22}a_{32}), & \beta_{33} &= -(a_{31}^2 + a_{32}^2 + a_{33}^2). \end{aligned} \tag{38}$$

Finally, we replace the elements of **B** in the system of input-output equations (34), (35), (36), and (37) by the relationships in (38) and estimate a_{ij} which ensures global concavity for the NQ function in (29). This replacement makes the system of input-output equations nonlinear in parameters a_{ij}.

For the NQ specification the price elasticities have the following expressions,

$$\eta_{ii} = \left[\frac{\beta_{ii} \sum_{j=1}^{n} \theta_j p_j - 2\theta_i \sum_{j=1}^{n} \beta_{ij} p_j + 2\theta_i^2 g(\mathbf{p})}{\left[\sum_{j=1}^{n} \theta_j p_j\right]^2} \right] \frac{p_i y}{x_i} \tag{39}$$

$$\eta_{ij} = \left[\frac{\beta_{ij} \sum_{j=1}^{n} \theta_j p_j - \theta_i \sum_{i=1}^{n} \beta_{ij} p_j - \theta_j \sum_{j=1}^{n} \beta_{ij} p_j + 2\theta_i \theta_j g(\mathbf{p})}{\left[\sum_{j=1}^{n} \theta_j p_j\right]^2} \right] \frac{p_j y}{x_j}. \tag{40}$$

4 Index Number Techniques

We also calculate the productivity growth in our sample industries using two widely used index number techniques—the Tornqvist index and the Fisher ideal index. Results from the index number techniques can be used to check the performance of the flexible functional forms. The Tornqvist index is the discrete approximation of Eq. (1), and the rate of technical change is calculated as follows

$$\ln y_t - \ln y_{t-1} - \sum_{i=1}^{n} \frac{1}{2}(s_{it} + s_{it-1})(\ln x_{it} - \ln x_{it-1}). \tag{41}$$

Equation (41) can also be written as

$$\ln \prod_{i=1}^{n} \left[\frac{(y/x_i)_t}{(y/x_i)_{t-1}} \right]^{\frac{1}{2}(s_{it} + s_{it-1})}. \tag{42}$$

The advantage of using the Tornqvist index is that it is exact for the linear homogeneous TL aggregator function (see Diewert 1976). However, in practice estimates of technical change obtained from the two approaches can be markedly different (Slade 1989).

With the Fisher ideal index, first the Fisher ideal quantity index for inputs is calculated as

$$I^t = \left[\frac{\sum_{j=1}^{n} p_j^t x_j^t}{\sum_{j=1}^{n} p_j^t x_j^{t-1}} \frac{\sum_{j=1}^{n} p_j^{t-1} x_j^t}{\sum_{j=1}^{n} p_j^{t-1} x_j^{t-1}} \right]^{1/2} \tag{43}$$

and then the quantity index for the single output is calculated as $Q^t = y^t / y^{t-1}$. Finally, the Fisher ideal total factor productivity index is computed as $(\frac{Q^t}{I^t} - 1)$. We also obtain a smoothed Fisher ideal index of total factor productivity, following Feng and Serletis (2008), by regressing the raw total factor productivity index on a constant and a time trend, and then calculating the fitted values.[6]

5 Data and Estimation Strategy

5.1 Data

In this study we use two different data sets that cover the period from 1961 to 2003 for the Canadian and U.S. manufacturing industries. Data for the Canadian manufacturing industries come from annual Canadian KLEMS database developed by the "Industry Multifactor Productivity Program" of Statistics Canada—see Baldwin et al. (2007) for a detailed description on the methodology used to construct this database.[7] Among the Canadian manufacturing industries we consider primary metal (NAICS 331), cement (NAICS 32731), and paper (NAICS 322) manufacturing industries in our sample.[8] The Canadian KLEMS data set contains annual information on chained Fisher quantity and price indexes for capital, labour, energy, material and services together with the information on quantity index of gross output as well as their nominal values.

The Jorgenson (2008) KLEM database, developed by Dale W. Jorgenson, and described in Jorgenson et al. (2000), provides the sample data for the U.S. manufacturing industries. The database is a combination of industry data collected from the U.S. Bureau of Labor Statistics (BLS) and the U.S. Bureau of Economic Analysis (BEA). It contains information on value and the price of output together with information on the values and prices for four inputs—capital, labour, energy, and material—for thirty-five U.S. industries covering the period from 1960 to 2005. The industries generally correspond to the 2-digit sectors in the Standard Industrial

[6]TFP estimates obtained from the smoothed Tornqvist index are almost identical to that obtained from the smoothed Fisher ideal index, and are not reported for brevity.

[7]Dissou and Ghazal (2010) utilize this dataset to examine energy substitutability in the primary metal and cement industries.

[8]The industries are at the L-level of aggregation in the North American Industry Classification System 2012.

Table 1 Input cost shares and growth rates of inputs, output and input prices: 1961–2003

	Canada			U.S.	
	Cement	Metal	Paper	Metal	Paper
Average annual growth rates					
Price of capital	5.80	4.63	2.10	3.62	2.14
Price of labour	5.30	5.42	5.45	4.42	5.13
Price of material	4.20	3.94	4.00	3.34	3.54
Price of energy	6.37	5.40	6.59	4.37	4.54
Quantity of capital	0.94	1.79	2.09	1.00	3.11
Quantity of labour	1.56	0.68	1.15	−0.52	0.81
Quantity of material	3.04	2.96	3.21	0.50	1.94
Quantity of energy	0.20	1.58	1.11	−0.13	1.15
Output	2.67	2.72	2.31	0.61	2.09
Average cost share					
Capital	17.54	9.03	16.08	9.52	13.29
Labour	24.38	20.16	22.67	20.86	25.55
Material	51.17	63.00	51.97	63.38	56.57
Energy	6.91	7.81	9.27	6.24	4.59

Classification (SIC) system.[9] In this study, we consider two of the thirty-five U.S. manufacturing industries—'primary metal' and 'paper and allied'—which roughly match with two of the Canadian industries in our sample and we also match the time period with the period covered in the Canadian KLEMS database. Average annual growth rates of inputs, output, and input prices along with the average input cost shares are presented in Table 1.

5.2 Estimation

For the TL specification, we perform joint estimation of the cost function (10) and the share equations (12) as a system of equations. Error disturbances, $\mathbf{v}_t$, which are assumed to have a multivariate normal distribution with zero mean and constant covariance over time, are added to the set of equations in the system. However, to avoid the problem of singularity we arbitrarily delete the material share equation. We use the iterative Zellner's technique for Seemingly Unrelated Regression (SUR) to estimate the system of equations. Parameter estimates of the deleted material share equation are obtained by using the linear homogeneity and symmetry restrictions.

[9] Young (2013), for example, uses this dataset to provide U.S. industry level estimates of the elasticity of substitution between labour and capital.

For the GL and the NQ specifications we only estimate the system of input-output equations (22) and (32), respectively, since the cost functions do not contain any additional parameters in both cases. All four input-output equations are used for the iterative SUR estimation. In estimating the NQ model for Canadian industries we normalize inputs prices in the first year to one. However, for the U.S. industries normalization is not required as input prices are equal to one for the base year (1996) in the data set.

If estimated models fail to satisfy the curvature condition, we follow the procedures explained in Sect. 3 to impose concavity. For the TL and the GL models, we follow the route suggested by Ryan and Wales (2000) to impose local concavity. It is important to note that the point of concavity imposition is arbitrary. If imposition of local concavity at all reference points fails to provide the expected concavity coverage at all sample points, then we choose the data point that provides the lowest number of curvature violations as the best approximation point and report results for that. For the NQ model we follow the technique described in Diewert and Wales (1987) to impose global concavity. For all three functional forms, imposing curvature conditions transforms the linear system of equations into nonlinear in parameters. Thus we use the nonlinear iterative SUR technique to estimate the systems restricted for concavity.

To verify whether the economic theoretical regularity conditions are satisfied by the estimated models we perform checks on positivity, monotonicity, and concavity. We evaluate fitted values of the cost function at each observation as a check for whether the estimated cost function is strictly positive. For monotonicity, we check whether the estimated input demand functions are all strictly positive at all data points. Necessary and sufficient condition for concavity is that the Hessian matrix of the cost function be negative semi-definite, and a real symmetric matrix will be negative semi-definite if it has non-positive eigenvalues. To check for the curvature condition, we compute eigenvalues of the estimated Hessian matrix of the cost function at each point in the sample space.

Although the full basic models are presented in Sect. 3, we test for the presence of technical change in all three functional forms using the likelihood ratio test. Since constant returns to scale are built in the datasets used in this study, test for the presence of returns to scale is ruled out. Moreover, in the TL model we test for the presence of neutral technical change. Finally, inclusion of the squared time trend is also tested for each model.

6 Result

6.1 Theoretical Regularity

Results from the likelihood ratio tests are presented in Table 2. Possibility of no technical change is rejected at 1% level of significance for all industries in the sample. In the TL model, the possibility of neutral technical change is also rejected

Table 2 Test statistics for likelihood ratio tests

	Canada			U.S.	
Functional form	Cement	Metal	Paper	Metal	Paper
No technical change					
TL	72.62	152.39	127.12	77.63	160.88
GL	124.84	185.23	179.20	101.04	248.86
NQ	138.10	160.68	200.64	105.23	206.22
No squared term of time					
TL	2.17*	42.13	8.71	33.91	28.64
GL	22.23	45.00	11.80*	66.53	97.00
NQ	29.76	44.68	10.85*	62.43	92.39
Neutral technical change					
TL	15.12	78.82	78.56	52.24	117.85

Notes: Test for the possibility of no technical change involves five restrictions in the TL, $\beta_t = \beta_{tt} = \beta_{it} = 0$, for $i = 1, \ldots, n-1$, and eight restrictions in both the GL and the NQ models, $\beta_{it} = \gamma_{it} = 0$, for $i = 1, \ldots, n$. Test for the possibility of neutral technical change in the TL imposes three restrictions, $\beta_{it} = 0$, for $i = 1, \ldots, n-1$. Again, test for the inclusion of squared time trend involves one parameter restriction in the TL, $\beta_{tt} = 0$, and in both the GL and the NQ models it involves four restrictions, $\gamma_{it} = 0$, for $i = 1, \ldots, n$. In all cases, $prob > \chi^2 = 0.0000$ except for the cases identified with *. For the Canadian Paper industry, in case of GL $prob > \chi^2 = 0.0189$, and in case of NQ $prob > \chi^2 = 0.0283$, while for the Canadian Cement industry, in TL $prob > \chi^2 = 0.1406$

for all industries. However, for the Canadian cement industry in the TL model and for the Canadian paper industry in the GL and the NQ models we fail to reject the null hypothesis of no squared time trend at 1% level of significance.

Results from estimating the TL, the GL, and the NQ models with and without curvature conditions reveal that the models satisfy positivity and monotonicity at all data points in the sample when curvature conditions are not imposed.[10] However, results for the concavity condition are not quite satisfactory. Counts on the incidence of curvature violations are reported in Table 3. When curvature conditions are not imposed only the GL model satisfies concavity at all data points for three out of the five sample industries; the TL and the NQ models violate curvature conditions in all cases.

Results from estimating the models with curvature conditions being imposed reveal that imposing curvature conditions does not always completely eliminate curvature violations in the TL and the GL models. For example, in the TL model for the Canadian metal and paper industries and in the GL model for the U.S. metal industry it fails to completely eliminate curvature violations. On the other hand, imposing global curvature conditions on the NQ reduces curvature violations to zero, as expected, in all cases.

[10]Tables with estimated coefficients and their standard errors are not reported here for brevity. However, they are available upon request to the corresponding author.

Table 3 Curvature violations by the estimated functional forms

	TL		GL		NQ	
	Curvature not imposed	Curvature imposed	Curvature not imposed	Curvature imposed	Curvature not imposed	Curvature Imposed
Canada						
Cement	11	0	43	0	43	0
Paper	43	3	0	—	43	0
Metal	15	3	0	—	43	0
U.S.						
Paper	43	0	0	—	43	0
Metal	28	0	43	1	43	0

Notes: Annual data, 1961–2003. Positivity and monotonicity violations are 0 in each case

Feng and Serletis (2008) find the performances of the TL and the GL models as poor when curvature conditions are being imposed on these models, and since both models fail to meet the curvature condition they do not provide productivity estimates based on these models. In this study, we present a more comprehensive case in the sense that our results include examples where imposing concavity on the TL and the GL models does successfully reduce curvature violations to zero. For example, in both the TL and the GL models for the Canadian cement industry it reduces curvature violations to zero. Moreover, the GL cost function satisfies all regularity conditions even without curvature being imposed for the Canadian metal and paper industries, for example. In what follows, we discuss the productivity and elasticity estimates only for those industries where the model satisfies all economic regularity conditions.

6.2 Productivity and Elasticity Estimates

Tables 4 and 5 report the average total factor productivity measures estimated with all three functional forms used in this paper. With the TL model, for the Canadian cement industry as well as for U.S. metal and paper industries, imposing curvature restrictions result in the satisfaction of all regularity conditions. However, the difference in TFP estimates, with and without curvature conditions, for these industries are noticeable in Table 4. Moreover, this discrepancy is also present in the estimates of autonomous and biased technical change. Given the importance of the rate and biases of technical change it is highly desirable that we get accurate estimates of those parameters. Results for the GL model in the case of Canadian cement industry, presented in Table 5, exhibit similar difference in TFP estimates with and without curvature being imposed. It can be seen more vividly in Fig. 1 which provides year-by-year TFP estimates for the Canadian cement industry estimated by the GL model with and without curvature being imposed.

Table 4 Average annual rates of technical change (%) using TL

	Autonomous technical change		Biased technical change		Total technical change	
	Curvature not imposed	Curvature imposed	Curvature not imposed	Curvature imposed	Curvature not imposed	Curvature imposed
Canada						
Cement	0.44	0.35	−0.06	0.02	0.38	0.37
U.S.						
Metal	0.17	0.18	−0.01	0.08	0.16	0.26
Paper	0.20	0.21	−0.01	0.07	0.19	0.28

Notes: Annual data, 1961–2003. Positivity and monotonicity violations are 0 in each case

Table 5 Average annual rates of technical change (%) using GL and NQ

	GL		NQ	
	Curvature not imposed	Curvature imposed	Curvature not imposed	Curvature imposed
Canada				
Cement	1.12	1.44	1.42	1.42
Paper	0.28	—	0.69	0.69
Metal	1.00	—	1.23	1.20
U.S.				
Paper	0.17	—	0.20	0.20
Metal	—	—	0.16	0.16

Notes: Annual data, 1961–2003. Positivity and monotonicity violations are 0 in each case

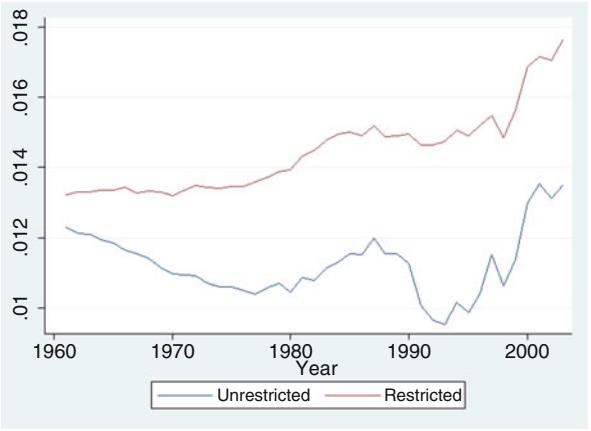


Fig. 1 Total factor productivity in Canadian cement industry using GL

The NQ model, on the other hand, performs better than the TL and the GL models when curvature condition is being imposed. Average TFP estimates for the NQ model are reported in Table 5. TFP estimates with curvature imposed are almost identical to the ones without curvature being imposed. Year-by-year TFP estimates

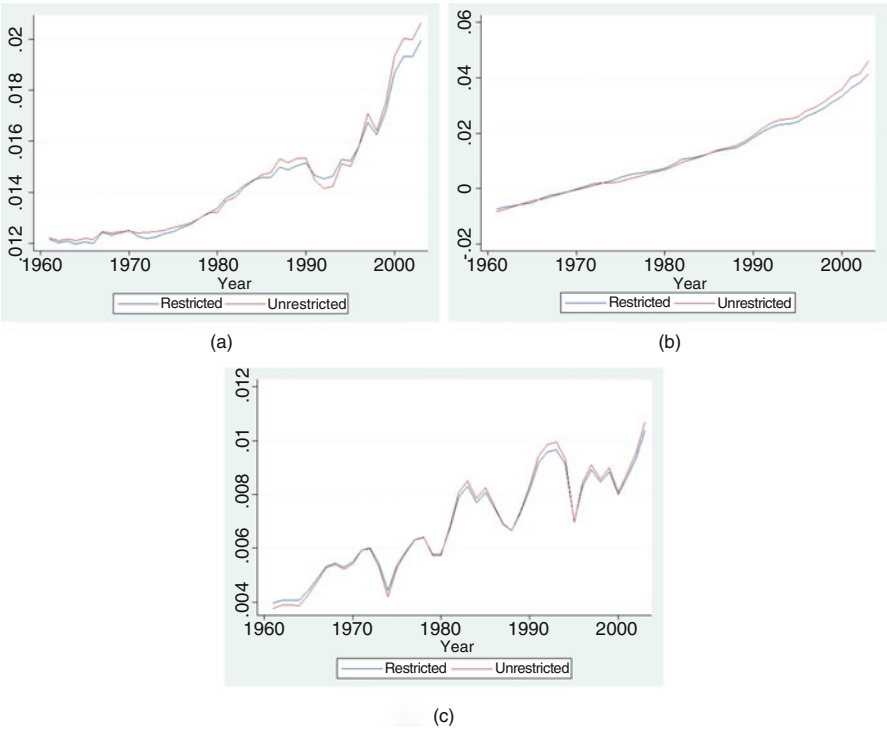


Fig. 2 Total factor productivity in Canadian manufacturing using NQ. (**a**) Cement. (**b**) Metal. (**c**) Paper

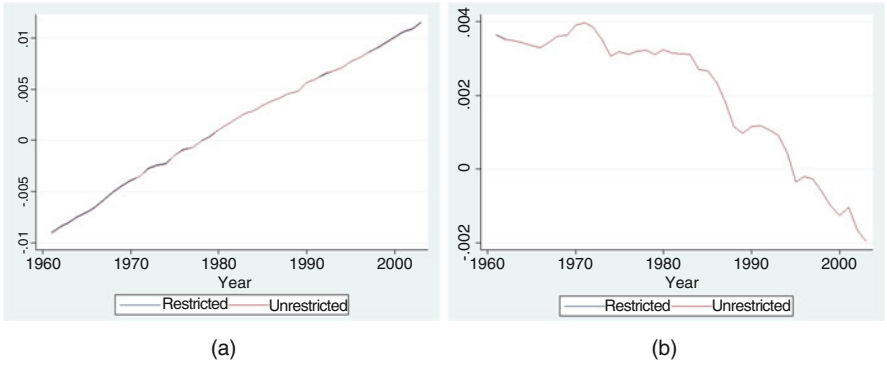


Fig. 3 Total factor productivity in U.S. manufacturing using NQ. (**a**) Metal. (**b**) Paper

for this model are presented in Figs. 2 and 3. It can be seen that the TFP estimates obtained from the restricted and unrestricted models are almost identical.

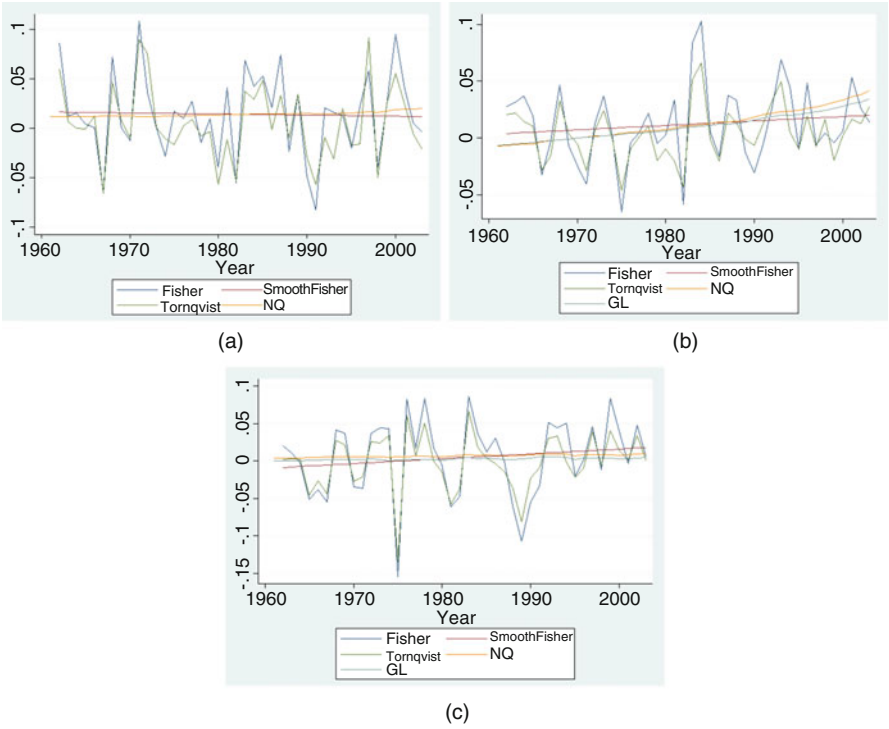


Fig. 4 Total factor productivity in Canadian manufacturing. (**a**) Cement. (**b**) Metal. (**c**) Paper

Figures 4 and 5 plot year-by-year TFP estimates from the NQ model, together with the productivity measures obtained from the Tornqvist, Fisher, and smoothed Fisher ideal indexes. Estimates from the GL model, when all economic regularity conditions are satisfied without curvature being imposed, are also included. The Tornqvist and the Fisher ideal indexes produce similar TFP estimates and for U.S. industries they are virtually identical. Productivity estimates from the NQ and GL models exhibit similar patterns during the sample period. Furthermore, both series of the GL and NQ estimates follow the smoothed Fisher ideal index very closely. Feng and Serletis (2008) use this criterion in evaluating the performance of NQ and AIM functional forms in providing TFP estimates. However, there is no theoretical correspondence of the smoothed Fisher Ideal index with the NQ or the AIM functional forms, although the correspondence between the Tornqvist index and the TL functional forms is well known (Diewert 1976). If we consider the benchmark adopted by Feng and Serletis (2008), both the NQ (with curvature being imposed) and the GL (without curvature being imposed) functional forms perform equally well in providing TFP estimates.

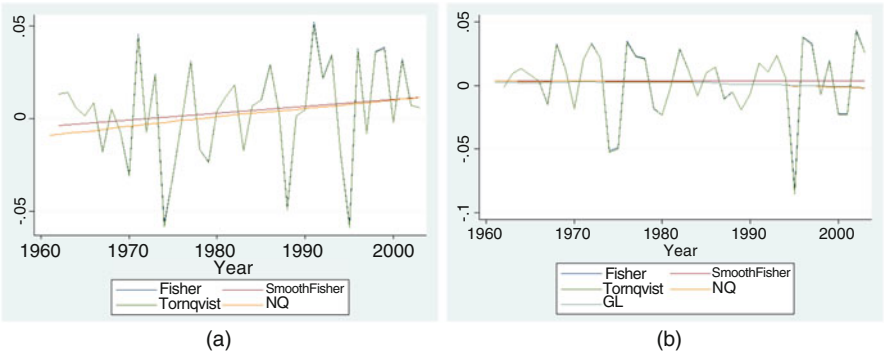


Fig. 5 Total factor productivity in U.S. manufacturing. (**a**) Metal. (**b**) Paper

Table 6 presents own and cross price elasticities of factor demand for our sample industries in U.S. and Canada. We calculate elasticities at the middle year of the sample period. Although not reported here for brevity, we also compute the elasticities when curvature conditions are not imposed. In most of the cases they follow closely to the ones reported in the table. In general, elasticity estimates obtained from the NQ model are smaller in absolute terms than the ones obtained from the TL and GL models. Estimated own price elasticities of demand in all cases have the correct sign. For all industries in the sample we find the derived factor demands inelastic as the estimated own price elasticities are below one in absolute terms.

7 Conclusion

In this paper we provide an empirical comparison and evaluation of three widely used locally flexible cost functional forms—the TL, the GL, and the NQ—in providing TFP estimates once they satisfy the economic theoretic regularity conditions. Estimation results for the sample industries provide us the opportunity to cover all possible cases that one might encounter in terms of curvature violations while estimating these three locally flexible functional forms: (i) curvature conditions are satisfied without curvature restrictions being imposed, (ii) curvature conditions are satisfied when curvature restrictions are imposed, and (iii) curvature conditions are not satisfied even with curvature restrictions being imposed. Findings reveal evidences of concavity violations even with curvature being imposed locally for the TL and the GL models. However, curvature violations reduce to zero for the NQ model when concavity is imposed globally.

When all economic regularity conditions are satisfied with curvature being imposed, our results suggest that the NQ model performs better than the other two models in providing TFP estimates. Estimates of productivity from the unrestricted

Table 6 Price elasticities for the Canadian and U.S. industries

	Canada				U.S.				
Factor i	Model	η_{Ki}	η_{Li}	η_{Mi}	η_{Ei}	η_{Ki}	η_{Li}	η_{Mi}	η_{Ei}
Metal									
(K)	NQ	−0.007	0.001	−0.006	0.002	−0.048	−0.136	0.158	−0.011
	GL	−0.190	−0.005	0.009	0.012	−0.557	0.016	0.046	−0.020
	TL	−0.364	−0.032	0.056	0.009	−0.362	0.069	0.039	−0.218
(L)	NQ	0.014	−0.124	−0.086	0.168	−0.122	−0.381	0.368	0.017
	GL	−0.085	−0.144	−0.010	0.157	0.077	−0.085	0.033	−0.015
	TL	−0.083	−0.435	0.092	0.440	0.404	−0.171	0.000	0.241
(M)	NQ	0.011	−0.028	−0.058	0.054	−0.091	0.487	−0.675	0.109
	GL	0.088	−0.007	−0.161	0.104	0.529	0.077	−0.088	0.049
	TL	0.437	0.279	−0.144	−0.023	0.497	0.001	−0.125	0.444
(E)	NQ	−0.030	0.151	0.150	−0.223	0.189	0.030	0.148	−0.114
	GL	0.186	0.156	0.162	−0.273	−0.048	−0.007	0.010	−0.013
	TL	0.010	0.188	−0.003	−0.426	−0.539	0.101	0.086	−0.466
Paper									
(K)	NQ	−0.011	−0.011	0.005	0.009	−0.168	−0.059	0.173	0.016
	GL	−0.037	−0.006	0.025	0.007	−0.281	−0.112	0.110	−0.058
	TL	−0.208	0.004	0.040	0.008	−0.189	−0.084	0.090	−0.058
(L)	NQ	−0.057	−0.144	−0.031	0.133	−0.071	−0.031	0.045	0.030
	GL	−0.013	−0.137	−0.004	0.173	−0.269	−0.422	0.227	0.112
	TL	0.009	−0.210	0.051	0.197	−0.320	−0.667	0.375	0.491
(M)	NQ	−0.017	−0.012	−0.020	0.014	−0.059	0.046	−0.328	0.077
	GL	0.037	−0.003	−0.080	0.050	0.585	0.505	−0.344	0.060
	TL	0.191	0.113	−0.143	0.248	0.568	0.619	−0.456	−0.057
(E)	NQ	0.056	0.166	0.046	−0.156	0.274	0.044	0.110	−0.123
	GL	0.012	0.146	0.059	−0.230	−0.036	0.029	0.007	−0.115
	TL	0.008	0.094	0.053	−0.452	−0.059	0.132	−0.009	−0.376
Cement									
(K)	NQ	−0.047	−0.036	−0.091	0.058	—			
	GL	−0.248	−0.020	−0.039	0.074				
	TL	−0.159	−0.048	0.086	−0.148				
(L)	NQ	−0.048	−0.264	0.346	0.067		—		
	GL	−0.034	−0.098	−0.035	0.055				
	TL	−0.112	−0.432	0.243	0.012				
(M)	NQ	−0.064	−0.169	−0.533	0.069			—	
	GL	−0.043	−0.022	−0.106	0.045				
	TL	0.396	0.475	−0.440	0.600				
(E)	NQ	−0.014	0.131	0.278	−0.194				—
	GL	0.326	0.139	0.180	−0.175				
	TL	−0.125	0.004	0.111	−0.465				

Notes: Annual data, 1961–2003. Elasticities are calculated at the middle year of the sample period

and the restricted NQ models are almost identical. However, the GL cost function performs equally well when all the regularity conditions are satisfied without curvature being imposed. Based on the evidences provided in this study we argue that since desired curvature coverage is not guaranteed in the TL and the GL models and local curvature imposition on these functional forms affects the productivity estimates, functional forms with global curvature conditions appear to be a better choice for econometric productivity estimation.

References

Acemoglu, D. (2002). Directed technical change. *Review of Economic Studies, 69*(4), 781–810.

Acemoglu, D. (2007). Equilibrium bias of technology. *Econometrica, 75*(5), 1371–1410.

Baldwin, J. R., Gu, W., & Yan, B. (2007). User guide for statistics Canada's annual multifactor productivity program. Canadian productivity review research paper, Statistics Canada, Catalogue no. 15–206–XIE.

Berndt, E. R. (1991). *The practice of econometrics: Classics and contemporary*. New York: Addison-Wesley Publishing Company.

Van Biesebroeck, J. (2007). Robustness of productivity estimates. *Journal of Industrial Economics, 55*(3), 529–569.

Binswanger, H. P. (1974a). A cost function approach to the measurement of elasticities of factor demand and elasticities of substitution. *American Journal of Agricultural Economics, 56*(2), 377–386.

Binswanger, H. P. (1974b). The measurement of technical change biases with many factors of production. *American Economic Review, 64*(6), 964–976.

Caves, D. W., Christensen, L. R., & Swanson, J. A. (1980). Productivity in US railroads, 1951–1974. *Bell Journal of Economics, 11*(1), 166–181.

Chalfant, J. A., & Wallace, N. E. (1992). Bayesian analysis and regularity conditions on flexible functional forms: Application to the US motor carrier industry. In W. E. Griffiths, H. Lutkepohl, & M. E. Bock (Eds.), *Readings in econometric theory and practice: A volume in honor of George Judge*. Amsterdam: North-Holland.

Christensen, L. R., Jorgenson, D. W., & Lau, L. J. (1971). Conjugate duality and the transcendental logarithmic production function. *Econometrica, 39*(4), 255–256.

Christensen, L. R., Jorgenson, D. W., & Lau, L. J. (1973). Transcendental logarithmic production frontiers. *Review of Economics and Statistics, 55*(1), 28–45.

Diewert, W. E. (1976). Exact and superlative index numbers. *Journal of Econometrics, 4*(2), 115–145.

Diewert, W. E., & Wales, T. J. (1987). Flexible functional forms and global curvature conditions. *Econometrica, 55*(1), 43–68.

Dissou, Y., & Ghazal, R. (2010). Energy substitutability in Canadian manufacturing econometric estimation with bootstrap confidence intervals. *Energy Journal, 31*(1), 121–148.

Feng, G., & Serletis, A. (2008). Productivity trends in US manufacturing: Evidence from the NQ and AIM cost functions. *Journal of Econometrics, 142*(1), 281–311.

Fisher, D., Fleissig, A. R., & Serletis, A. (2001). An empirical comparison of flexible demand system functional forms. *Journal of Applied Econometrics, 16*(1), 59–80.

Gallant, A. R., & Golub, G. H. (1984). Imposing curvature restrictions on flexible functional forms. *Journal of Econometrics, 26*(3), 295–321.

Geweke, J. (1986). Exact inference in the inequality constrained normal linear regression model. *Journal of Applied Econometrics, 1*(2), 127–141.

Griffiths, W. E., O'Donnell, C. J., & Cruz, A. T. (2000). Imposing regularity conditions on a system of cost and factor share equations. *Australian Journal of Agricultural and Resource Economics, 44*(1), 107–127.

Hulten, C. R. (2001). Total factor productivity: A short biography. In C. R. Hulten, E. R. Dean, & M. J. Harper (Eds.), *New developments in productivity analysis* (pp. 1–54). University of Chicago Press. http://www.nber.org/chapters/c10122

Jaffe, A. B., Newell, R. G., & Stavins, R. N. (2003). Technological change and the environment. *Handbook of Environmental Economics, 1*, 461–516.

Jin, H., & Jorgenson, D. W. (2010). Econometric modeling of technical change. *Journal of Econometrics, 157*(2), 205–219.

Jorgenson, D. W. (2008). 35 Sector KLEM. Harvard Dataverse, V1, http://hdl.handle.net/1902.1/10684

Jorgenson, D. W., Stiroh, K. J., Gordon, R. J., & Sichel, D. E. (2000). Raising the speed limit: US economic growth in the information age. Brookings papers on economic activity (pp. 125–235).

Kohli, U. (1992). Production, foreign trade, and global curvature conditions: Switzerland, 1948–1988. *Swiss Journal of Economics and Statistics, 128*(1), 3–20.

Lau, L. J. (1978). Testing and imposing monotonicity, convexity and quasiconvexity constraints. In M. Fuss & D. McFadden (Eds.), *Production economics: A dual approach to theory and applications* (Vol. 1, pp. 409–453). Amsterdam: North Holland.

León-Ledesma, M. A., McAdam, P., & Willman, A. (2010). Identifying the elasticity of substitution with biased technical change. *American Economic Review, 100*(4), 1330–1357.

Morey, E. R. (1986). An introduction to checking, testing, and imposing curvature properties: The true function and the estimated function. *Canadian Journal of Economics, 19*(2), 207–235.

Moschini, G. (1999). Imposing local curvature conditions in flexible demand systems. *Journal of Business and Economic Statistics, 17*(4), 487–490.

Ryan, D. L., & Wales, T. J. (1998). A simple method for imposing local curvature in some flexible consumer-demand systems. *Journal of Business and Economic Statistics, 16*(3), 331–338.

Ryan, D. L., & Wales, T. J. (2000). Imposing local concavity in the translog and generalized Leontief cost functions. *Economics Letters, 67*(3), 253–260.

Slade, M. E. (1989). Modelling stochastic and cyclical components of technical change: An application of the Kalman filter. *Journal of Econometrics, 41*(3), 363–383.

Terrell, D. (1996). Incorporating monotonicity and concavity conditions in flexible functional forms. *Journal of Applied Econometrics, 11*(2), 179–194.

Wiley, D. E., Schmidt, W. H., & Bramble, W. J. (1973). Studies of a class of covariance structure models. *Journal of the American Statistical Association, 68*(342), 317–323.

Young, A. T. (2013). US elasticities of substitution and factor augmentation at the industry level. *Macroeconomic Dynamics, 17*(04), 861–897.

Productivity Growth, Poverty Reduction and Income Inequality: New Empirical Evidence

Mahamat Hamit-Haggar and Malick Souare

Abstract There is a long-standing view that economic growth is the most powerful instrument for reducing poverty. In dynamic economies most economic growth comes from productivity growth, and yet the literature concerning the relationship between productivity changes and poverty is sparse. Against this backdrop, this paper examines the impact of productivity growth on income and human poverty, and assesses the role played by the income distribution in that relationship. Using cross-country data to conduct a regional comparative analysis, we find that productivity growth is more relevant for poverty reduction than the more commonly used indicator economic growth – a finding that is robust across regions. We also find that the poverty-reducing impact of productivity growth is stronger in countries with relatively low income inequality. These findings suggest that countries attempting to reach their objectives of eradicating poverty should pursue policies that foster productivity growth; and that productivity growth that is accompanied by progressive distributional change is even better for alleviating poverty.

Keywords Productivity Growth · Economic Growth · Poverty · Inequality · Regional Analysis

JEL Classification: E24, I32, O15, O47

The views expressed in this paper are those of the authors and do not necessarily represent those of Innovation, Science and Economic Development Canada or the Government of Canada.

M. Hamit-Haggar (✉)
CERDI, Université Clermont Auvergne, Clermont-Ferrand, France
e-mail: mahamat.hamit-haggar@etu.uca.fr

M. Souare
Innovation, Science and Economic Development Canada, Ottawa, ON, Canada
e-mail: malick.souare@canada.ca

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_11

1 Introduction

According to the Organisation for Economic Co-operation and Development – OECD (2008), if you asked a typical person to list the major problems that the world faces today, the likelihood is that "poverty and inequality" would be one of the first things they mentioned. For example, as shown in Fig. 1, in the most recent World Values Survey, 65.3 percent of respondents across the world stated that "people living in poverty and need" was the most serious problem facing the world, and 60 percent considered it as such in their own country.[1] Reducing poverty represents a key objective and a fundamental challenge for policymakers in both developing and developed countries. It has received a greater attention since the adoption of the Millennium Development Goals (MDGs) in 2000, which among other things, targeted the halving of the rate of poverty between 1990 and 2015. This attention was recently renewed by the establishment of a new set of Sustainable Development Goals (SDGs) in September 2015. The SDGs replace the MDGs and shift the poverty reduction objective from halving to eliminating by 2030. What is more, unlike the MDGs that focused solely on developing countries, the SDGs are a *universal* framework for achieving sustainable development outcomes in *all* countries; i.e., they apply to all countries, including developed ones.

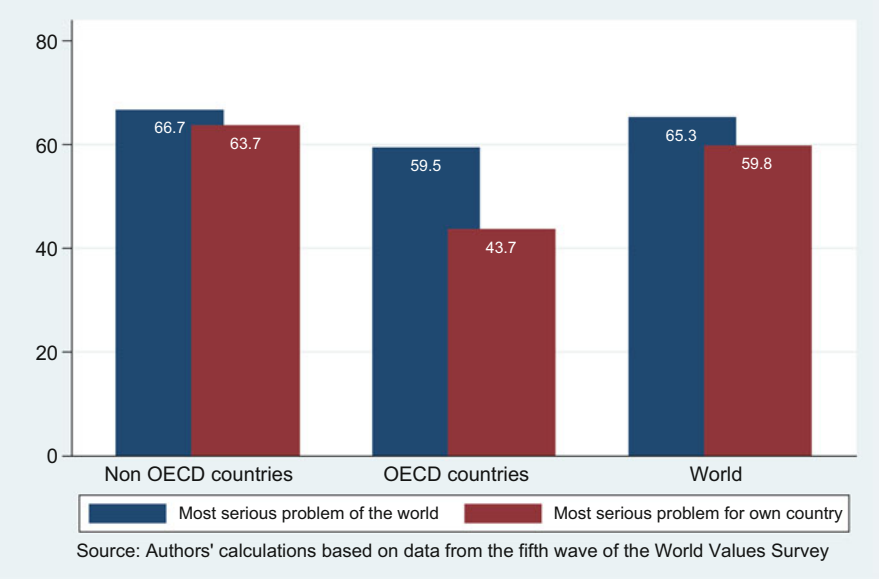


Fig. 1 Percentage of people who consider "living in poverty and need" as the most serious problem for the world or for the own country

[1] World Values Survey – Wave 5 (2005–2009), Online Data Analysis.

There is a long-standing view that economic growth is a prerequisite for poverty reduction. The World Bank (2006) reported that countries that have historically experienced the greatest reduction in poverty are those that have experienced prolonged periods of sustained economic growth. For example, over the 1981–2000 period, China's poverty rate fell from more than 50 percent to about 8 percent, thanks to an impressive per capita growth rate of almost 8.5 percent a year. Similarly, between 1993 and 2002 Vietnam cut its poverty rate in half, from about 58 percent to about 29 percent, by growing at almost 6 percent a year. Besides, several recent studies have also found that the higher is income inequality within a country, the more limited is the impact of growth on reducing poverty. Thus, economic growth that is associated with progressive and redistributive policies will reduce poverty more than growth that leaves the distribution unchanged.

Nonetheless, in dynamic economies, most of the economic growth comes from productivity growth. Relatedly, there is a strong consensus that productivity is the single most important determinant of a nation's living standard or its level of real income over the medium to long run. To put it differently, productivity sets the sustainable level and path of prosperity that a country can achieve. From this perspective, productivity growth appears to be the key for attaining the global objective of eradicating poverty and improving living standards. Yet, poverty reduction strategies seldom focus explicitly on productivity.

An interesting illustration of the prominence of productivity improvements (for poverty reduction and rising living standards) stems from the Africa's recent experience. According to the International Monetary Fund – IMF (2015), the African continent has enjoyed a strong and persistent economic growth for more than a decade. For 15 years GDP growth rates have averaged over 5 percent. However, the World Economic Forum – WEF (2013 and 2015) underscored this impressive and unprecedented growth on the continent has not translated into any meaningful poverty reduction or rapidly improving living standards, as has happened in other regions (such as much of emerging and developing Asia) with a similar growth performance. In other words, more than a decade of consistently high growth rates have not yet trickled down to significant parts of the population.[2] Indeed, as potential underlying causes, the WEF has recurrently argued that *low and falling productivity figures* in Africa are at the core of these differences in living standards relative to other regions – see the next section for some cross-regional stylized facts.

Thus, it seems that the poverty-reducing (or the living standard enhancing) impact of economic growth depends more on the source of growth as opposed to growth itself. For example, for many observers (see, Lipton, 2012) there is no doubt that the decisive underlying driver of African growth performance was the high commodity prices. However, an increase in commodity prices does not necessarily

[2]Actually, nearly one out of two Africans (i.e., about 47 percent) continue to live in extreme poverty – this figure is measured against a threshold of US$1.25 dollar a day. See http://povertydata.worldbank.org/poverty/region/SSA for details.

translate into higher productivity unless it is accompanied by appropriate measures and policies (WEF, 2011).[3] Thus, it is important to give a special attention to productivity among the sources of economic growth.[4]

Nonetheless, in other developing regions (such as several countries in Latin America and the Caribbean) that have experienced little poverty reduction or even increasing poverty despite some economic growth, high and growing income inequality has been identified as a major culprit. Thus, it has been suggested that while economic growth is important for poverty reduction, growth that is accompanied by inequality-reducing policies is even more suitable.

Although there is an abundant literature on the links among economic growth, poverty and inequality, very few studies have investigated the relationship of productivity growth and income distribution to poverty reduction. Against this backdrop, this paper contributes to the literature by analysing empirically the relationships among these variables. To this end, the paper uses cross-country data to conduct a regional comparative analysis. The rest of the paper is organized as follows. In the next section, we present some stylized facts and a brief review of the literature. These stylized facts underscore the importance of productivity being a key driver of economic growth for poverty reduction and improvement in living standards; and the literature review presents briefly the few existing studies that explored the link between productivity (growth) and poverty, which typically overlooked the role of income distribution. Section 3 outlines the empirical framework, describes the variables and data sources, and presents empirical results for the relationship between productivity growth and poverty, and that between productivity growth, income distribution and poverty. These results are contrast with the corresponding relationships involving economic growth. Finally, section 4 discuses some policy implications and concludes the paper.

[3] As another case in point, over the 2000s, Canada fared relatively well in terms of economic growth as the country posted the largest growth rate among the G7 countries. However, many academics and policymakers have expressed concerns about future prospects for Canada's economic growth and improvements in living standards, mainly because (over the same period) the country has suffered from a stubborn lack of productivity growth relative to the U.S. and other OECD countries. A consensus from this hotly policy debate is that lower Canadian productivity is the main cause of the country's lower living standard compared with the U.S., and closing this productivity gap is the only sustainable way to reduce the two-country income gap.

[4] One reason that may explain why productivity has been neglected in the literature as a determinant of poverty reduction is that economic growth already subsumes productivity growth. It may have been felt that its impact was already covered. Moreover, the difficulty of obtaining reliable labour input data in most developing countries, needed to calculate labour productivity, may have contributed to the use of GDP per capita or mean income in poverty reduction studies as well.

2 Some Stylised Facts and Review of Literature

2.1 The Importance of Productivity as a Key Driver of Economic Growth

We begin this section by stressing the role of productivity enhancements as a driver of economic growth for poverty reduction and improvement in living standards. To this end, we compare Africa's performance with other regions, such as developing Asia, to help understand that a fast-growing but generally low-productivity economy does not offer strong prospects for poverty reduction and rising living standards. Although high and persistent economic growth rates characterize Africa for more than a decade – growth rates have averaged well above 5 percent in the past 15 years, the continent has not experienced the rapidly improving living standards that have been seen in other regions with a similar growth performance (see Fig. 2).

As shown in Fig. 2, although both Africa and Southeast Asia had approximately the same levels of GDP per capita in the 1990s, Southeast Asia's GDP per capita has since risen considerably more rapidly than sub-Saharan Africa. According to several WEF's reports on the Africa competiveness (published on a biennial basis since 1998), low and falling productivity fundamentals are at the core of these differences in living standards.

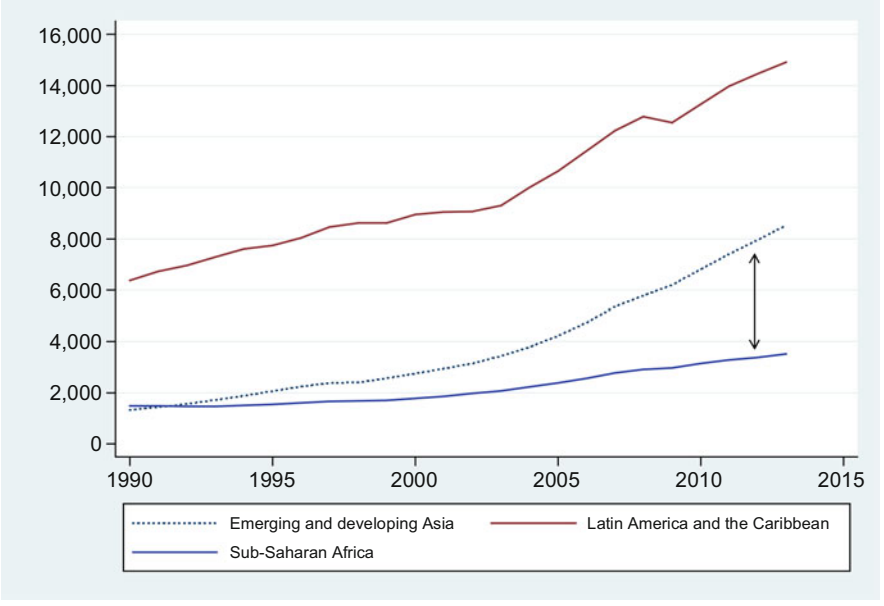


Fig. 2 Prosperity and economic growth, 1990–2013-GDP based on purchasing power parity (PPP) per capita, current int'l dollars (Source: WEF, 2015.)

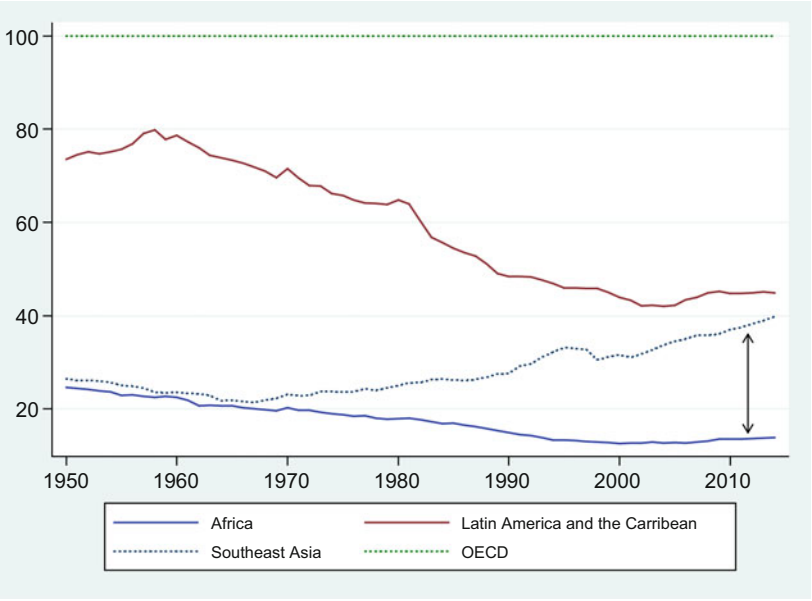


Fig. 3 Cross-regional productivity Labor productivity per person employed in 1990 US$ (converted at Geary Khamis PPPs) (Source: The Conference Board Total Economy Database™, June 2015, http://www.conference-board.org/data/economydatabase. Note: **Southeast Asia** includes Cambodia, Indonesia, Malaysia, Myanmar, Philippines, Singapore, Thailand, and Vietnam; **Africa** includes Algeria, Angola, Burkina Faso, Cameroon, Côte d'Ivoire, the Democratic Republic of Congo, Egypt, Ethiopia, Ghana, Kenya, Madagascar, Malawi, Mali, Morocco, Mozambique, Niger, Nigeria, Senegal, South Africa, Sudan, Tanzania, Tunisia, Uganda, Zambia, and Zimbabwe; **Latin America and the Caribbean** includes Argentina, Barbados, Bolivia, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Guatemala, Jamaica, Mexico, Peru, St. Lucia, Trinidad and Tobago, Uruguay, and Venezuela; **OECD countries** include Australia, Austria, Belgium, Canada, Chile, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Rep., Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom, and the United States)

Figure 3 compares labor productivity – as a proxy for overall productivity – in Africa with that of other regions for the past 50 years. Although Africa and Southeast Asia started from similar, very low levels, labor in Southeast Asia has since become more productive, effectively converging toward the OECD average. In contrast, as Fig. 3 shows, not only has Africa been trailing Southeast Asia, but in fact the productivity gap between the two regions deepened between 1960 and 2005. What is more, using data at the sectoral levels, WEF (2015) shows that across sectors – from agriculture to manufacturing and services – productivity levels remain low in Africa. This illustrates the importance of growth being driven by productivity enhancements that are associated with rising living standards.

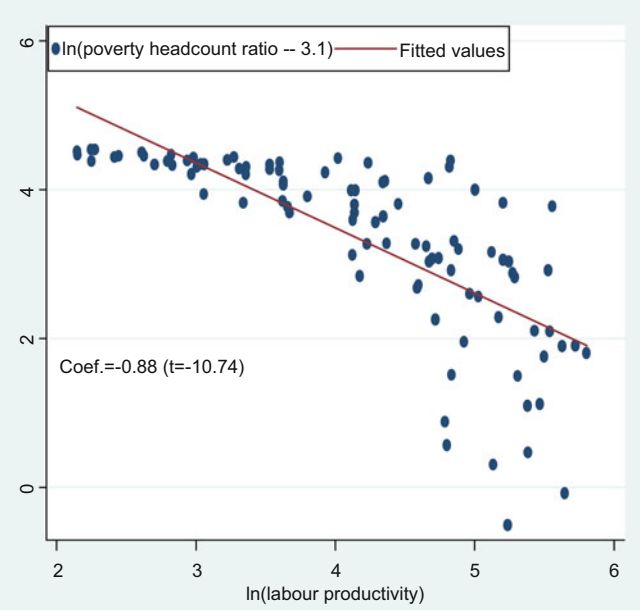


Fig. 4 Poverty and productivity growth

Finally, as a preliminary investigation of the relative relevance of productivity and economic growth for poverty reduction, Figs. 4 and 5 present the scatterplot diagrams of the headcount poverty measure (vertical axis) and logged labour productivity and real GDP (horizontal axis), respectively. It clearly appears that the negative relationship is more pronounced with productivity growth compared to that with the economic growth, and also with the former relationship showing a far better goodness of fit.

2.2 A Brief Review of Literature

The literature concerning with the relationship between productivity and poverty is sparse. As noted above, although economic growth dominates world talks on poverty, the specific role of productivity (which is the most important source of long-term economic growth) is often overlooked and poverty reduction strategies rarely underscore the importance of productivity. Moreover, the few existing studies exploring the link between productivity and poverty also typically ignore the role of income distribution. We now briefly review the few empirical literature on productivity (growth) and poverty.

Pineau (2004) argues that in a world with limited available capital, especially for developing countries, and rapid population growth, productivity increases are

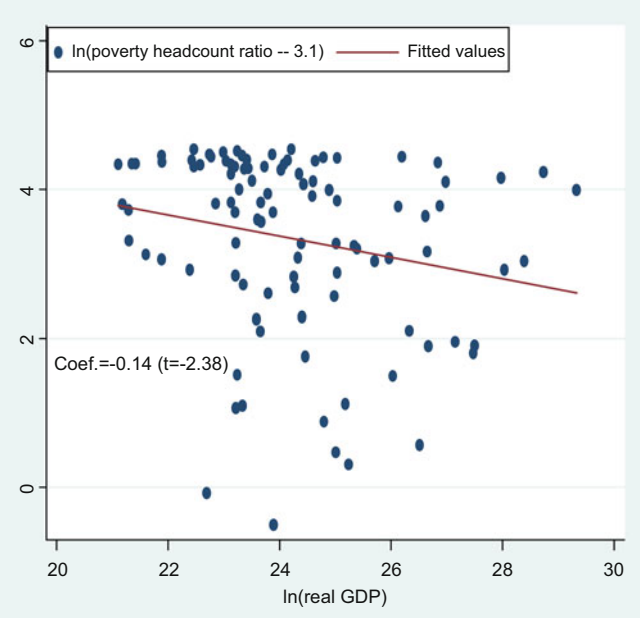


Fig. 5 Poverty and economic growth

the only source of growth that can lead to a sustainable expansion of income per capita. He claims that in the long run and at the aggregate level, other sources of growth cannot result in significant per capita increases, as the additional output from these sources is proportional to additional inputs, which expand mostly with population growth. This leaves per capita income unchanged, unless higher productivity is achieved. In a Peruvian case study, Pineau documents the potential of a specific micro-level institution to increase the productivity of labour, leading to a significant poverty reduction across many dimensions (including material well-being, psychological well-being, access to basic infrastructure, and capacity to manage assets). He concludes that productivity is indeed what poor people need to get out of poverty.

Fluet and Lefebvre (1987) contend that increased productivity produces, among other things, an easier access to material goods and services (through higher incomes and/or lower prices)[5] and therefore reduces poverty. Relatedly, they investigate how (total factor) productivity improvements in Canadian manufacturing were apportioned among labour, capital, materials and government through an increase in the price of these factors or through an increase in taxes levied on factor inputs, and consumers through a decrease in the industry selling prices. They find

[5]Productivity gains will either lead to rising factor prices or to a reduction in the price of output. In the first case, workers and capital owners gain, while in the second case, consumers gain.

that about half of the manufacturing productivity gains were passed on to the rest of the economy through lower relative prices. And the share of labour price increases accounted for the bulk of the other half.

In a United States study, Hayes *et al*. (1995) examines the relationship between labour productivity growth and poverty rate, accounting for other macroeconomic variables. The authors do not assume that the relationship is unidirectional from productivity to poverty. Specifically, their hypothesis is that there may be bidirectional causality between poverty and changes in productivity. The empirical results suggest that feedback does exist between productivity and poverty – i.e. productivity growth both affects and is affected by changes in poverty.[6] Thus, as a public policy implication of their results, they propose that policies or measures intended to affect productivity growth or poverty must be designed simultaneously. For example, education and training programs may affect productivity not only directly but also indirectly by lowering future poverty rates and raising the level of human capital in the future. Besides, using data on developing countries, CSLS (2003) makes a strong case for productivity increases as a tool to reduce poverty.

As it is common in developing countries that most of the poor live in rural areas, and that a large share of them depend on agriculture for a living and spend large shares of their income on food, many studies have focused on the specific link between agricultural productivity growth and poverty. For example, using data on the Indian economy, Datt and Ravallion (1998) investigate the impact of farm productivity (as measured by output per acre) on rural poverty. They find that in the short run, higher agricultural productivity reduces poverty through expanded employment opportunities or more abundant harvests. However, in the long run, increasing agricultural productivity reduces poverty through higher wages and lower relative food prices. Byerlee *et al.* (2009) review 12 country case studies and use bivariate analysis to compare agricultural growth per worker across countries. They show that the countries with the highest agricultural growth per worker experienced the greatest rate of rural poverty reduction. Fan *et al.* (1999) measure the relationship between total factor productivity and poverty outcomes by investigating returns on different productivity increasing investments. They find that investments in roads, agricultural research, development, and extension had the greatest impact on both productivity and poverty reduction in India. Using data on Ethiopia and Madagascar, Abro *et al.* (2014) and Minten and Barrett (2008) find respectively that agricultural productivity growth holds the key to poverty reduction. Other empirical studies also reveal that productivity growth in agriculture helps reduce poverty (Cervantes-Godoy & Dewbre, 2010; Christiaensen *et al.*, 2010; Irz *et al.*, 2001; Majid, 2004; Ravallion & Chen, 2005; Thirtle *et al.*, 2003; World Bank, 2008). Thus, regardless of which productivity measure is used (output per worker, output per unit of land, or TFP), empirical studies support the idea that improvements in agricultural productivity are important for poverty reduction.

[6]In other words, poverty, through low investment in human capital, e.g., also reduces labour productivity growth.

To sum up, it emerges that despite the relative lack of literature on productivity and poverty, the few existing studies show that rising productivity does contribute to poverty reduction. Thus, the relationship between the two variables is an important one and further research on the subject could make the fight against poverty more effective. As a result, this paper contributes to the literature by not only investigating the relationship between the two variables across many regions, but also accounting for the role income inequality plays in that relationship.

3 Empirical Framework, Data Description and Empirical Results

3.1 Productivity Growth and Poverty

As mentioned earlier, the literature has focused on the relationship between economic growth and poverty and not on the relationship between productivity growth and poverty. This paper contributes to the literature by attempting to close this knowledge gap. In order to investigate the impact of productivity growth on the incidence of poverty and compare its antipoverty effectiveness with that of economic growth, we start by estimating the following two basic specifications:

$$lnPov_{it} = \theta_i + \beta_1 lnProd_{it} + \varepsilon_{it} \tag{1}$$

$$lnPov_{it} = \theta_i + \alpha_1 lnGDP_{it} + \omega_{it} \tag{2}$$

where *Pov* represents either the income or human-based poverty measures, *Prod* denotes productivity (as proxied by labour productivity), *GDP* is real gross domestic product, i and t stand respectively for country and year, θ_i represents country fixed or random effects, ε and ω are error terms. Because of the fixed/random effect term, it is noteworthy that only countries with data on poverty, labour productivity and real income for more than 1 year will be included in the regressions (see Table 6). Besides, with the double logarithmic equations, the parameters of interest, β_1 and α_1, are respectively the productivity and economic growth elasticity of poverty, and therefore a key magnitude in assessing the antipoverty effectiveness of the former variables.[7] Following the development literature, they correspondingly represent a measure of the "poverty efficiency" of productivity and economic growth. We expect that $\widehat{\beta}_1 < 0$ and $\widehat{\alpha}_1 < 0$, but more importantly $\left|\widehat{\beta}_1\right| > |\widehat{\alpha}_1|$ – that is, the poverty-reducing impact of productivity growth should be greater than that of economic growth.

[7]Moreover, one other advantage of the use of the logarithmic form is that it weakens the potential heteroscedasticity problems, in addition to reducing the impact of outlier observations.

As mentioned above, the poverty reduction target with the MDGs was to halve, between 1990 and 2015, the proportion of people living in poverty. However, the SDGs, adopted in 2015, have a more ambitious agenda, seeking to eliminate rather than reduce poverty by 2030. Interestingly, using Eq. (1), e.g., one can derive the annual rate of productivity growth that will allow achieving a particular poverty reduction goal between any given two periods, $t1$ and $t2$, as follows:

$$g_{LP} = \frac{\ln\left(Pov_{t2}/Pov_{t1}\right)}{\beta_1{}^{*}\left(t2 - t1\right)}, \quad \forall Pov_{t2} \; and \; Pov_{t1} \neq 0 \tag{3}$$

where Pov_{t2} and Pov_{t1} represent the poverty measure in year $t2$ and $t1$, respectively.[8]

At this stage, we consider three measures of poverty, namely the poverty headcount ratio at \$1.9/day and at \$3.1/day poverty line, as well as the multidimensional poverty index (MPI). The first two are the traditional income-based poverty measures, indicating the proportion of people living with less than a threshold level of income per day. With the \$1.9 poverty line, the focus is on the poorest of the poor, while with the \$3.1 poverty line, the poverty concept is more inclusive. The third measure of poverty (i.e., the MPI) is based on a human development notion and uses a number of factors to determine poverty beyond income-based criterion. The MPI provides a comprehensive picture of people living in poverty by using information from 10 indicators which are organized into three dimensions: education (years of schooling and school attendance), health (child mortality and nutrition), and living standard (electricity, sanitation, water, floor, cooking fuel and assets).[9] – see Table 7 for more details on the construction of the MPI. Table 8 provides a complete list of the definitions, data sources and summary statistics of the regression variables, including the later ones – these statistics pertain to the (regional) pooled country-year sample.

We first estimate Eqs. (1) and (2) using the entire sample across all countries and employing panel data techniques of fixed- and random-effects models (see Table 1). Table 1 shows the results for the three variables of poverty described above. It clearly emerges that regardless of the poverty measures and estimation methods used, the poverty-reducing efficiency of productivity growth is *greater* than that of economic growth – the elasticity of poverty with respect to productivity growth is 1.4 to 3 times higher in absolute terms. Moreover, the goodness of fit, as measured by the R-squared coefficients, for the relationship between poverty and productivity growth is 9 to 11 times higher as well. In other words, productivity growth seems to

[8]Note that equation (3) is not defined if $Pov_{2030} = 0$ (i.e., at the target of the SDGs in 2030), and therefore it will be used only for illustrative purposes for targets around that goal. Moreover, using Eq. (2), one may derive a similar annual rate for economic growth.

[9]Moreover, the MPI reflects *both* the ***incidence*** or headcount ratio (H) of poverty – the proportion of the population that is multidimensionally poor – and the average ***intensity*** (A) of their poverty – the average proportion of indicators in which poor people are deprived. Thus, the MPI is calculated by multiplying the incidence of poverty by the average intensity across the poor ($H \times A$). A person is identified as poor if he or she is deprived in at least one third of the weighted indicators.

Table 1 Productivity and economic growth and poverty – whole sample estimation

	Poverty headcount ratio – 3.1		Poverty headcount ratio – 1.9		Multidimensional poverty index	
Dependent variable	Productivity growth	Economic growth	Productivity growth	Economic growth	Productivity growth	Economic growth
Fixed-effects model	−1.931*** (0.111)	−1.347*** (0.088)	−2.403*** (0.139)	−1.734*** (0.106)	−1.238*** (0.246)	−0.836*** (0.174)
Observations	702	726	686	707	152	152
R2	0.471	0.044	0.489	0.053	0.619	0.064
Random-effects model	−1.240*** (0.067)	−0.588*** (0.057)	−1.657*** (0.084)	−0.862*** (0.075)	−1.232*** (0.104)	−0.413***(0.097)
Observations	702	726	686	707	152	152
R2	0.471	0.044	0.489	0.053	0.619	0.064
Hausman test (FE vs. RE)	Chi2(1) = 60.96 (prob = 0.000)	Chi2(1) = 130.86 (prob = 0.000)	Chi2(1) = 45.89 (prob = 0.000)	Chi2(1) = 132.47 (prob = 0.000)	Chi2(1) = 0.00 (prob = 0.981)	Chi2(1) = 8.60 (prob = 0.003)

The standard errors are reported in parentheses beside the parameters estimates. ***, **, * denote statistical significance at the 1%, 5% and 10% levels, respectively.

have a more explanatory power in terms of poverty reduction than economic growth. Overall, the results suggest that productivity growth is a more important driver of reductions in poverty than economic growth. Besides, for each poverty variable we run a Hausman test to choose between fixed- and random-effects models. For this test, random effects (RE) is preferred under the null hypothesis due to higher efficiency, while under the alternative fixed effects (FE) is at least consistent and thus preferred. As shown in Table 1, for all cases except one, the use of fixed-effects model is recommended.

As the poverty concept is more inclusive with $3.1 poverty line and that there is limited data availability for the multidimensional poverty measure across regions, we use poverty measure based on $3.1 a day to estimate Eqs. (1) and (2) across regions[10] – see Table 2. This table only reports results as they pertain to the suitable models suggested by the subsequent Hausman test. Once again, one can observe that across all six regions, the marginal poverty-reducing impact of productivity growth is greater than that of economic growth, with the largest gap occurring in Sub-Saharan Africa.[11] Moreover, as for the whole sample estimation above, a stronger fit of the relationship between poverty and productivity growth (relative to that between poverty and economic growth) is also found across all regions. These results show again the importance of giving much attention to (labor) productivity growth as a poverty-reducing variable relative to economic growth.

The bottom part of Table 2 displays the annual rate of productivity growth that would be needed to reduce poverty rate by 80% in the world or in that region between 2015 and 2030 if productivity growth alone is to reduce poverty. Comparing this to the historical productivity growth rate between 1990 and 2012 (shown in the last row of Table 2), we see that at the average historical pace only two regions in the world, East Asia and pacific and South Asia, will succeed in meeting the above target of poverty reduction by 2030. Putting it differently, these are the two developing regions where the historical rate of productivity growth exceeds the rate of growth needed to reduce poverty by 80% over the next 15 years. In the remaining regions, the historical growth rate is considerably below that required to cut poverty rate 80%. It is also noteworthy that Sub-Saharan Africa experienced the lowest historical productivity growth rate.

Table 3 provides across regions both the poverty trends (proportion and absolute numbers of people living in poverty) and the decomposition of economic growth into its components (i.e., population growth, growth in the employment to total population ratio, and labour productivity growth) with their percentage contribution. From Panel A of Table 3, we see that in East Asia, e.g., the poverty rate drops by about 62 percentage points, from 83.9% in 1990 to 22.3% in 2012, and numbers in

[10]Nonetheless, the results in the whole sample estimation (Table 1) indicate that the poverty-reducing power of both productivity and economic growth are lower when the poverty line is set higher (i.e., at $3.1, relative to $1.9).

[11]This distinguished evidence on the African continent may be explained by the fact that a largest proportion of the poor depend on labor income for a living.

Table 2 Productivity and economic growth and poverty – regional estimation (Dep. variable: poverty $3.1 a day)

	Whole sample	East Asia and Pacific	Eastern Europe and Central Asia	Latin America and Caribbean	Middle East and North Africa	South Asia	Sub-Saharan Africa
Elasticity of poverty with respect to productivity growth	−1.931*** (0.111)	−2.521*** (0.357)	−1.068** (0.133)	−1.618*** (0.292)	−0.929*** (0.334)	−2.900*** (0.226)	−2.050*** (0.218)
R2	0.471	0.484	0.62	0.607	0.811	0.413	0.453
Elasticity of poverty with respect to economic growth	−1.347*** (0.088)	−2.287*** (0.361)	−0.835*** (0.080)	−1.515*** (0.216)	−0.646*** (0.231)	−2.344*** (0.393)	−1.245*** (0.157)
R2	0.044	0.117	0.021	0.033	0.12	0.056	0.068
Annual rate of productivity growth needed to reduce poverty 80% by 2030	5.56%	4.26%	10.05%	6.63%	11.55%	3.70%	5.23%
Historical productivity growth 1990–2012	2.35%	7.61%	2.52%	1.35%	2.15%	4.83%	1.23%

The standard errors are reported in parentheses beside the parameters estimates. ***, **, * denote statistical significance at the 1%, 5% and 10% levels, respectively.

Table 3 Poverty trends and the decomposition of economic growth into its components

Panel A: Poverty trends								
	Poverty rate (% below $3.10)				Number of poor (1,000,000)			
Regions	1990	1999	2008	2012	1990	1999	2008	2012
East Asia & Pacific	83.86	65.00	35.33	22.23	1345	1168	684	443
Europe & Central Asia	7.85	19.56	7.32	6.24	19	48	19	16
Latin America & Caribbean	30.83	26.20	15.38	11.96	115	113	75	61
Middle East & North Africa	24.90	20.27	16.03	na	56	55	51	na
South Asia	81.76	na	67.88	54.5	926	na	1074	913
Sub-Saharan Africa	76.11	77.87	70.71	66.97	387	506	584	617
World*	53.90	31.32	36.81	28.93	2847	1890	2488	2051

Panel B: GDP growth and its components				
	GDP Growth	Pop	Empl/Pop	GDP/worker
East Asia & Pacific	8.28	0.97	−0.35	7.61
	100.00	11.66	−4.17	91.97
Europe & Central Asia	2.61	0.40	−0.31	2.52
	100.00	15.39	−11.90	96.67
Latin America & Caribbean	3.22	1.47	0.37	1.35
	100.00	45.54	11.62	41.99
Middle East & North Africa	4.08	1.89	0.00	2.15
	100.00	46.34	0.06	52.70
South Asia	6.25	1.77	−0.41	4.83
	100.00	28.38	−6.50	77.32
Sub-Saharan Africa	4.09	2.74	0.08	1.23
	100.00	67.01	1.94	30.14

poverty fall from over 1.3 billion to 443 million. Over the same period, Sub-Saharan Africa registered the lowest reduction in poverty rate, 9.1 percentage points – from 76.1% to 67%. In addition, this modest reduction is accompanied with an increase in the absolute numbers of people living in poverty, from 387 in 1990 to 617 million in 2012. Interestingly, "putting these trends in perspective" with the magnitude of the labor productivity's contribution to economic growth (as shown in Panel B of Table 3), we observe that the labor productivity growth contributed 92% to economic growth in East Asia, compared with only 30% in Sub-Saharan Africa, which represents the lowest contribution across regions. Thus, it emerges that countries that have historically experienced the greatest reduction in poverty are those in which labour productivity growth made the largest contribution to economic growth.

This evidence helps underline the centrality of productivity growth for poverty reduction efforts. Thus, understanding and implementing what drives productivity growth is key to achieving the Sustainable Poverty Reduction Goals. Nonetheless, as shown above the required productivity growth to reduce poverty rate 80% by 2030, e.g., is large relative to historical averages in many regions. This suggests that productivity growth by itself seems unlikely to be enough to attain the Sustainable

Development Goals for poverty reduction in those regions. Thus, an optimal poverty reduction strategy should also identify policies and factors that can directly reduce poverty, even if productivity growth does not increase, or which can improve the mapping of productivity growth onto poverty (i.e., increase the elasticity of poverty with respect to productivity growth). To this end, we explore the role of income redistribution, along with productivity, in the next sub-section.

3.2 Productivity Growth, Income Distribution and Poverty

In the literature on poverty, alongside economic growth progressive income redistribution has been considered as one of the main avenues leading to poverty reduction. There are two main reasons for this. One is that, in general, for a fixed level of income, progressive distributional change will shift resources from the richer to the poorer and thus lead to poverty reduction[12] – this is a one-shot instant impact on poverty resulting from the pure positive redistribution effect. The other reason is that poverty is more responsive to growth the more equal the income distribution. In other words, reducing income inequality improves the mapping of growth onto poverty by increasing (in absolute value) the growth elasticity of poverty and hence makes future growth more effective in reducing poverty. In sum, an improvement in the distribution of income has a double poverty-reducing impact. Therefore, we add a measure of inequality in income distribution (*Ineq*) into Eqs. (1) and (2).

$$lnPov_{it} = \theta_i + \beta_1 lnProd_{it} + \beta_2 Ineq_{it} + \varepsilon_{it} \tag{4}$$

$$lnPov_{it} = \theta_i + \alpha_1 lnGDP_{it} + \alpha_2 Ineq_{it} + \omega_{it} \tag{5}$$

where *Ineq* is measured by the log of the Gini index or by the standard deviation of the income distribution in logs.

Table 4 shows the results when Eqs (4) and (5) are estimated using the whole sample and the three variables of poverty described above. The first observation is that increases in income inequality are positively associated with increases in poverty, regardless of the inequality indicator. The second observation is that the introduction of income inequality dampens the impacts of both productivity and economic growth on poverty (in absolute terms).[13] The third noteworthy observation is that productivity growth is still more efficient in reducing poverty

[12]The exception is when per capita income levels are below the poverty line, in which case progressive distributional change leads to increasing poverty.

[13]The negative impact of higher income inequality on the poverty-reducing impact of both productivity and economic growth also emerges when we break the entire sample into high and low income inequality countries.

Table 4 Productivity and economic growth, distribution and poverty – whole sample estimation

	Poverty headcount ratio – 3.1		Poverty headcount ratio – 1.9		Multidimensional poverty index	
Panel A						
Elasticity of poverty with respect to productivity growth	−1.707*** (0.107)	−1.796*** (0.111)	−2.136*** (0.128)	−2.185*** (0.133)	−2.211* (1.168)	−2.564* (1.263)
Log of Gini Index	2.498*** (0.258)		3.632*** (0.311)		1.716 (1.843)	
Standard deviation of the income distribution in logs		3.295*** (0.565)		5.731*** (0.653)		2.017 (5.592)
R2	0.612	0.551	0.669	0.625	0.551	0.491
Panel B						
Elasticity of poverty with respect to economic growth	−1.175*** (0.085)	−1.234*** (0.088)	−1.522*** (0.099)	−1.548*** (0.104)	−1.433 (0.902)	−1.785 (1.076)
Log of Gini Index	2.607*** (0.266)		3.654*** (0.321)		1.961 (1.911)	
Standard deviation of the income distribution in logs		3.243*** (0.589)		5.491*** (0.683)		1.705 (6.395)
R2	0.108	0.066	0.126	0.089	0.082	0.039

The standard errors are reported in parentheses beside the parameters estimates. ***, **, * denote statistical significance at the 1%, 5% and 10% levels, respectively.

than economic growth, irrespective of poverty measures. Moreover, the equations including productivity growth also still have a more explanatory power in terms of poverty reduction. Together, these findings suggest that productivity growth is more relevant for poverty reduction than the more commonly used economic growth, and that productivity growth that is accompanied by progressive income distributional change is even better for alleviating poverty.

Table 5 presents the results across regions when we use the poverty measure based on $3.1 a day and consider the Gini index as a measure of inequality. The three key findings just discussed above (from all sample estimation in Table 4) still hold across each region. More importantly, in all cases, productivity growth has both a larger impact on poverty reduction and a greater explanatory power. Besides, all these findings are robust to measuring inequality by the standard deviation of the income distribution in logs (see Table 9 in appendix).

4 Alternative Measures of Poverty Covering Developed Countries

Viewed from a developed country perspective, living on $1.9 or even $3.1 a day is unthinkable. This notion of poverty, generally used in developing countries, is referred to in the literature as the absolute poverty measure. However, applying this measure of poverty to most developed countries would result in virtually nobody being classified as poor. The literature on cross-country trends of poverty in developed countries is largely based on the relative poverty concept, generally defined as the proportion of individuals with disposable income less than 50% of the median income in a given country. Thus, there is not a one-to-one relationship between economic/productivity growth and relative poverty, as the former may reduce or increase the latter.[14] This explains why some people criticize the concept of relative poverty on the grounds that it is to do with 'inequality' rather than 'poverty' (see Nielsen, 2009). Relatedly, referring to the measurement of poverty when the poverty line is set as a function of the income distribution, Fields (1980) writes that "this is more an inequality measure than a poverty measure, because if everyone's income were to increase by the same percentage, poverty would be unaffected." Similarly, in Duclos and Makdissi (2007, 2004) relative poverty is

[14]For example, suppose an economic or productivity growth that increases real incomes along the entire income distribution. If one can expect that growth to reduce absolute poverty, the effect on relative poverty will vary depending on which income strata benefits the most. Imagine, e.g., that those with mid-level incomes experienced a stronger real income growth than those with low incomes, this would move more people below the relative poverty line, although their real incomes actually increased.

Table 5 Productivity and economic growth, distribution and poverty – regional estimation (Dep. variable: poverty $3.1 a day)

	Whole sample	East Asia and Pacific	Eastern Europe and Central Asia	Latin America and Caribbean	Middle East and North Africa	South Asia	Sub-Saharan Africa
Panel A							
Elasticity of poverty with respect to productivity growth	−1.707*** (0.107)	−2.128*** (0.364)	−1.039*** (0.119)	−1.484*** (0.272)	−0.922*** (0.331)	−2.588*** (0.293)	−1.755*** (0.212)
Log of Gini Index	2.498*** (0.258)	2.496*** (0.850)	1.921*** (0.309)	4.227*** (0.781)	0.466 (0.359)	1.368 (0.842)	3.177*** (0.568)
R2	0.617	0.636	0.777	0.564	0.817	0.490	0.607
Panel B							
Elasticity of poverty with respect to economic growth	−1.175*** (0.085)	−1.893*** (0.369)	−0.814*** (0.073)	−1.356*** (0.207)	−0.612** (0.236)	−1.440*** (0.530)	−1.050*** (0.150)
Log of Gini Index	2.607*** (0.266)	2.506*** (0.865)	1.637*** (0.282)	4.057*** (0.782)	0.317 (0.369)	3.444** (1.441)	3.506*** (0.563)
R2	0.108	0.195	0.027	0.119	0.134	0.099	0.174

The standard errors are reported in parentheses beside the parameters estimates. ***, **, * denote statistical significance at the 1%, 5% and 10% levels, respectively.

depicted as a restricted or censored inequality measure.[15] In other words, relative poverty is considered as a special way of measuring income inequality. Thus, this feature of the relative concept of poverty makes it less suitable to use in the study of trends in poverty.

5 Conclusions and Policy Implications

The United Nations has set a universal goal to eradicate world's poverty by 2030. To achieve that challenging goal, strong and sustained economic growth is rightfully considered as the main driving force behind such a pace of poverty reduction. In dynamic economies, however, most of the economic growth comes from productivity growth. From this perspective, productivity growth is then the key for attaining this global objective. Nonetheless, the literature concerning the relationship between productivity changes and poverty is very sparse. Against this background, this paper examines the impact of productivity growth on income and human poverty, and assesses the role played by the income distribution in that relationship.

Using cross-country data to conduct a regional comparative analysis, the paper finds that productivity growth is more relevant for poverty reduction than the more commonly used indicator economic growth – a finding that is robust across all studied regions. The paper shows that countries that have historically experienced the greatest reduction in poverty are those in which labour productivity growth made the largest contribution to economic growth. Thus, although not always recognized in the literature on poverty, these findings indicate that productivity should be acknowledged as being central to poverty reduction, and hence be at the center of any potential successful strategy aimed at reducing poverty. The paper also finds that the level of income inequality plays an important role in the relationship between the two variables. The poverty-reducing impact of productivity growth is stronger in countries with relatively low income inequality. Thus, productivity growth that is accompanied by progressive distributional change is even better for alleviating poverty.

However, even though this paper provides strong support for the view that productivity growth is essential for poverty reduction and should be a priority for each and every country, it is much harder to identify and implement appropriate policies that will increase productivity growth. Many determinants drive the productivity of an economy, and the literature has stressed the following policies and factors underpinning national productivity growth: policies and actions that support well-

[15]Nonetheless, commenting on the links between poverty and (relative) inequality, Sen (1983) argues that poverty is an absolute notion in the space of capabilities (i.e., the set of functionings available to an individual), but that it often takes a relative form in the space of income or consumption because the achievement of some social functionings requires more income in a richer society.

functioning institutions (public and private), extensive and efficient infrastructure investments (in transport, telecommunications, and energy, e.g.), a stable macroeconomic environment, human capital development (through quality education and training, and better healthcare), openness and competition, structural transformation of the economy (i.e., the transfer of resources from low productivity sectors to high productivity sectors), the adoption and use of existing new technologies (which are typically embodied in new machinery and equipment, including ICT capital) and investment in technological innovation.

Nonetheless, in line with well-known economic theory of stages of development, although all of the pro-productivity policies described above matter to a certain extent for all economies, they affect different countries in different ways, depending on their stage of development. More specifically, the best way for Guinea, e.g., to improve its productivity is not the same as the best way for Canada to do so – this is because the two countries are in different stages of development. Less-advanced countries can substantially enhance their productivity by improving institutions, building infrastructure, or reducing macroeconomic instability, but all these factors eventually run into diminishing returns as countries move along the development path. For more advanced economies, these 'basic' productivity-enhancing factors are no longer sufficient for meaningfully increasing productivity; they have to increasingly rely on technological innovation. In sum, the relative importance of each productivity driver depends on a country's particular stage of development.

Moreover, as the paper shows that the level of income inequality mediates the relationship between productivity growth and poverty, with a progressive income distribution increasing the impact of productivity on poverty, some focus on inequality reduction is not unreasonable. Nonetheless, it is also noteworthy that unlike in advanced economies, the potential for achieving redistribution via conventional tax and transfer systems is limited in developing countries. Development practitioners have instead advocated for other measures such as increasing access to credit, strengthening property rights and improving the delivery of public services. Relatedly, some attention also needs to be paid to the distributional impact of productivity growth, which in turn suggests a focus on specific drivers of productivity growth that can directly benefit the poor. The bottom line is that policies geared toward alleviating poverty must include strategies for sustainable productivity growth along with those aimed to improve income distribution.

Acknowledgements We would like to thank participants in the 2016 Canadian Economics Association Conference and North American Productivity Workshop, as well as the anonymous referee for their valuable comments and suggestions.

Annexes

Table 6 List of countries by regions

East Asia & Pacific	East Europe & Central Asia	Latin America & Caribbean	Middle East & North Africa	South Asia	Sub-Saharan Africa
Cambodia	Albania	Belize	Djibouti	Bangladesh	Angola
China	Armenia	Bolivia	Iran, Islamic Rep.	**Bhutan**	Benin
Fiji	Azerbaijan	**Brazil**	Jordan	India	Botswana
Indonesia	Belarus	Colombia	Morocco	Maldives	Burkina Faso
Lao PDR	Bosnia and Herzegovina	**Costa Rica**	Tunisia	Nepal	Burundi
Malaysia	Bulgaria	Dominican Republic	West Bank and Gaza	Pakistan	Cabo Verde
Mongolia	Georgia	**Ecuador**		Sri Lanka	Cameroon
Papua New Guinea	Kazakhstan	Guatemala			Central African Republic
Philippines	Kosovo	Guyana			Chad
Thailand	Kyrgyz Republic	**Haiti**			Congo, Dem. Rep.
Timor-Leste	Macedonia, FYR	Honduras			Congo, Rep.
Vietnam	Moldova	Jamaica			Cote d'Ivoire
	Montenegro	**Mexico**			**Ethiopia**
	Romania	Nicaragua			Gambia, The
	Serbia	Panama			Ghana
	Tajikistan	Paraguay			Guinea
	Turkey	Peru			Guinea-Bissau
	Turkmenistan				Kenya
	Ukraine				Lesotho
	Uzbekistan				Madagascar
					Malawi
					Mali
					Mauritania
					Mauritius
					Mozambique
					Namibia
					Niger
					Nigeria
					Rwanda
					Senegal
					Sierra Leone
					South Africa
					Swaziland

(continued)

Table 6 (continued)

East Asia & Pacific	East Europe & Central Asia	Latin America & Caribbean	Middle East & North Africa	South Asia	Sub-Saharan Africa
					Tanzania
					Togo
					Uganda
					Zambia
In bold are countries that are not included when estimating the MPI model					

Table 7 The dimensions, indicators, deprivation cut-offs and weights of the MPI

Dimensions of poverty	Indicator	Deprived if...	Weight
Education	Years of Schooling	No household member has completed 5 years of schooling.	1/6
	Child School Attendance	Any school-aged child is not attending school up to class 8.	1/6
Health	Child Mortality	Any child has died in the family.	1/6
	Nutrition	Any adult or child for whom there is nutritional information is malnourished.	1/6
Living Standard	Electricity	The household has no electricity.	1/8
	Improved Sanitation	The household's sanitation facility is not improved (according to MDG guidelines), or it is improved but shared with other households.	1/8
	Improved Drinking Water	The household does not have access to improved drinking water (according to MDG guidelines) or safe drinking water is more than a 30-minute walk from home, roundtrip.	1/8
	Flooring	The household has a dirt, sand or dung floor.	1/8
	Cooking Fuel	The household cooks with dung, wood or charcoal.	1/8
	Assets ownership	The household does not own more than one radio, TV, telephone, bike, motorbike or refrigerator and does not own a car or truck.	1/8

Table 8 Definitions and sources of data

Variable	Definition	Mean	Std. Dev.
Pov	Poverty headcount ratio at $ 1.90 a day % population (constant 2011 PPP $)	19.84	21.64
	Poverty headcount ratio at $ 3.10 a day % population (constant 2011 PPP $)	35.16	27.17
	Multidimensional poverty index	37.65	29.95
Ineq	Log(GINI index)	3.74	0.22
	Standard deviation of Income distribution in log	0.37	0.09
Prod	Productivity defined as real GDP per person employed (constant 2011 PPP $)	97.68	86.29
GDP	Real GDP (constant 2011 PPP $))	5164.16	4122.65

Source: World Development Indicators

Table 9 Productivity and economic growth, distribution and poverty – regional estimation (Dep. variable: poverty $3.1 a day)

	Whole sample	East Asia and Pacific	Eastern Europe and Central Asia	Latin America and Caribbean	Middle East and North Africa	South Asia	Sub-Saharan Africa
Panel A							
Elasticity of poverty with respect to productivity growth	−1.797*** (0.111)	−2.446*** (0.373)	−0.989*** (0.115)	−1.533*** (0.285)	−0.894** (0.333)	−2.591*** (0.286)	−1.880*** (0.222)
Standard deviation of the income distribution in logs	3.296*** (0.565)	1.117 (1.539)	4.601*** (0.643)	6.685*** (1.899)	1.160 (0.922)	4.484* (2.628)	3.817*** (1.214)
R2	0.552	0.517	0.781	0.574	0.823	0.499	0.545
Panel B							
Elasticity of poverty with respect to economic growth	−1.235*** (0.089)	−2.209*** (0.378)	−0.773*** (0.072)	−1.439*** (0.214)	−0.599*** (0.239)	−1.505*** (0.535)	−1.092*** (0.163)
Standard deviation of the income distribution in logs	3.243*** (0.590)	1.117 (1.553)	3.945*** (0.598)	6.566*** (1.916)	0.804 (0.957)	10.257** (4.660)	3.944*** (1.272)
R2	0.066	0.128	0.024	0.072	0.122	0.095	0.104

The standard errors are reported in parentheses beside the parameters estimates. ***, **, * denote statistical significance at the 1%, 5% and 10% levels, respectively.

References

Abro, Z. A., Alemu, B. A., & Hanjra, M. A. (2014). Policies for agricultural productivity growth and poverty reduction in rural Ethiopia. *World Development, 59*, 461–474.

Byerlee, D., Diao, X., & Jackson, C. (2009). *Agriculture, rural development, and pro-poor growth: Country experiences in the post- reform Era*. https://doi.org/10.1146/annurev.resource.050708.144239.

Center for the Study of Living Standards (CSLS). (2003). *Productivity Growth and Poverty Reduction in Developing Countries*. Background Paper prepared for the 2004 World Employment Report of the International Labour Organization, CSLS Research Report 2003–06, Ottawa.

Cervantes-Godoy, D., & Dewbre, J. (2010). *Economic importance of agriculture for poverty reduction*. OECD Food, Agriculture and Fisheries Working Papers No. 23, OECD Publishing.

Christiaensen, L., Demery, L., & Kuhl, J. (2010). The (evolving) role of agriculture in poverty reduction: An empirical perspective. *Journal of Development Economics, 96*, 239–254.

Datt, G., & Ravallion, M. (1998). Farm productivity and rural poverty in India. *Journal of Development Studies, 34*, 62–85.

Duclos, J.-Y., & Makdissi, P. (2004). Restricted and unrestricted dominance for welfare, inequality and poverty orderings. *Journal of Public Economic Theory, 6*(1), 145–164.

Duclos, J.-Y., & Makdissi, P. (2007). Restricted inequality and relative poverty. In J. Bishop & Y. Amiel (Eds.), *Inequality and poverty: papers from the society for the study of economic inequality's inaugural meeting, research on economic inequality* (Vol. 14, pp. 255–280). Elsevier.

Fan, S., Hazell, P., & Thorat, S. (1999). Linkages between government spending, growth, and poverty in rural India. *(Research Report No. 100).*. Washington, D.C.: International Food Policy Research Institute. Retrieved from http://www.ifpri.org/sites/default/files/publications/rr110.pdf.

Fields, G. (1980). *Poverty, inequality and development*. Cambridge: Cambridge University Press.

Fluet, C., & Lefebvre, P. (1987). The Sharing of total factor productivity gains in Canadian manufacturing: A price accounting approach. 1965-1980. *Applied Economics, 19*, 245–257.

Hayes, K. J., Slottje, D. J., Nieswiadomy, M. L., & Wolff, E. N. (1995). The relationship between productivity changes and poverty in the United States. *Journal of Income Distribution, 4*, 107–119.

IMF. (2015). *Regional economic outlook: Sub-Saharan Africa: Navigating headwinds*. Washington, DC: IMF. Available at http://www.imf.org/external/pubs/ft/reo/2015/afr/eng/pdf/sreo0415.pdf.

Irz, X., Lin, L., Thirtle, C., & Wiggins, S. (2001). Agricultural productivity growth and poverty alleviation. *Development Policy Review, 19*, 449–466.

Lipton, M. (2012). *Income from Work: The food-population-resource crisis in the 'short Africa'*. Medford, MA: Leontief Prize Lecture, Tufts University. Available at http://www.ase.tufts.edu/gdae/about_us/leontief/LiptonLeontiefPrizeComments.pdf.

Majid, M. (2004). Employment strategy papers: Employment analysis unit employment strategy department. In *Reaching millennium goals: How well does agricultural productivity growth reduce poverty*. Geneva, Switzerland: International Labor Organization (ILO).

Minten, B., & Barrett, C. B. (2008). Agricultural technology, productivity, and poverty in madagascar. *World Development, 36*, 797–822.

Nielsen, L. (2009). *Global relative poverty*. IMF working paper No. 93. Washington, D.C.: International Monetary Fund.

OECD. (2008). *Growing unequal? Income distribution and poverty in OECD Countries*. Paris: France.

Pineau, P. O. (2004). Productivity to reduce poverty: Study of a micro-level institution in Peru. *International Productivity Monitor*, 62–75.

Ravallion, M., & Chen, S. (2005). China's (uneven) progress against poverty. *Journal of Development Economics, 82*, 1–42.

Sen, A. K. (1983). Poor, relatively speaking. *Oxford Economic Papers, 35*, 153–169.
Thirtle, C., Lin, L., & Piesse, J. (2003). The impact of research-led agricultural productivity growth on poverty reduction in Africa, Asia and Latin America. *World Development, 31*, 1959–1975.
World Bank. (2006). Poverty reduction and growth: virtuous and vicious circles. Written by Guillermo E. Perry, Omar S. Arias, J. Humberto Lopez, William F. Maloney and Luis Serven.World Bank Latin American and Caribbean studies,
World Bank. (2008). World development report: Agriculture for development. Washington, DC: The World Bank.
World Economic Forum. (2011). The Global Competitiveness Report 2011–2012. Geneva: World Economic Forum.
World Economic Forum. (2013). The Africa Competitiveness Report 2013. Geneva: World Economic Forum.
World Economic Forum. (2015). The Africa Competitiveness Report 2013. Geneva: World Economic Forum.

The Contribution of Productivity and Price Change to Farm-level Profitability: A Dual Approach Analysis of Crop Production in Norway

Habtamu Alem

Abstract Previous studies estimating TFP and its components can be criticized for not considering unobserved heterogeneity in their model. Moreover, the studies focused on the technical evaluation of a sector. However, the technical evaluation alone reveals how well farmers use the physical production process. There is a need to closely examine the cost efficiency of the farmers. In this study, we used a cost function (dual) approach to facilitating the decomposition and estimation of TFP components. Using a translog stochastic cost function, we estimated the level and source of productivity and profitability change for crop producing family firms in Norway. We used the true random effect to account for farm heterogeneity. The analysis is based on 23 years unbalanced panel data (1991–2013) from 455 only crop-producing firms with a total of 3885 observations. The result indicates that average annual productivity growth rate in grain and forage production was – 0.11% per annum during the period 1991–2013. The profit change was −0.14% per annum.

Keywords Productivity · Profit · Panel data · Crop production and cost function

JEL Classification: C23, D24, M21

1 Introduction

Increasing agricultural productivity to feed the growing population is contemporary development challenge for developing and developed countries. Compared to other European countries, the total acreage of agricultural land in Norway is small and, because of the topography, many fields are scattered and often steep. These factors make agriculture costly. In recognition of these conditions,

H. Alem (✉)
Norwegian Institute of Bioeconomy Research and Norwegian University of Life Science, As, Norway
e-mail: habtamu.alem@nibio.no

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_12

the Norwegian government has assigned relatively large subsidies to the agriculture sector compared with other countries. The main goal of the Norwegian government is sustainable agricultural production in all regions. Thus, livestock production is a common practice all over the country. However, eastern Norway and central regions are with geographical, soil and climatic conditions relatively favorable for grain and forage production.

Agricultural productivity growth in Norway is a topic of continuing interest to researchers and policy makers who aim to improve economic sustainability in the sector. The Norwegian government white paper report no. 9 (2011–2012) stated that the main goal of the Norwegian agriculture sector is to increase food production to keep up the present level self-sufficiency. There is a need to measure and evaluate the economic performance of farms to suggest possible improvements to achieve agricultural policies.

The economic performance of a firm can be measured by the efficiency and productivity measures. Efficiency estimation involves estimating the frontier based on production, cost or profit functions and measuring the performance of the farmers to the frontier (Coelli et al. 2005). The word productivity[1] in economics is a broad concept, but this study focused on total factor productivity (TFP) as an appropriate measure of productivity. TFP is the ratio of aggregate output to aggregate inputs, which shows how much output firms produce from a given quantity of inputs. The dynamics of TFP can be measured by the evolution of the TFP over time. TFP change is a widespread quantitative economic instrument used to evaluate the performance and sustainability of agricultural systems over time. It has proven valuable for policy measures geared towards fostering agricultural development (Melfou et al. 2007).

Few studies conducted on the performance of agricultural production in Norway particularly focused on a dairy farm. For instance (Koesling et al. 2008; Kumbhakar et al. 2012; Lien et al. 2010; Odeck 2007; Sipilainen et al. 2013). We still very little known about the performance of the Norwegian agricultural sector First, the previous studies ignored forage production in spite of it being major output in the Norwegian agriculture with, for instance, 2400 mill.kg of forage produced in the year 2013 (Statistics Norway 2016). Second, the analysis for this study is based on extensive farm-level panel data set for a long period of observations (1991–2013). The firms, in the long run, can change all inputs and allows choosing the combination of inputs that reduce the cost of production at a given output. Moreover, previous productivity studies failed to consider unobserved heterogeneity within the regions or groups. The efficiency estimated in the previous models didn't distinguish individual heterogeneity from the inefficiency. In these models, all the time-invariant heterogeneity is confounded into inefficiency. Thus, the inefficiency component might be picking up heterogeneity in addition to or even instead of inefficiency (Greene 2005).

[1]Productivity is the ratio of output per unit of input(s) so that it can be measured in different forms. For instance a partial productivity measure uses only one input e.g. productivity of labor = aggregate output/labor.

The rest of the paper organized as follows. Section two presents the theoretical framework with a detailed derivation of productivity and profitability change components from the cost function. Section three describes the empirical model while section four discusses the data and definition of variables used in the cost function. Empirical estimation and results presented in section five. The final section encompasses a summary of our findings and conclusions.

2 Theoretical Framework

2.1 Theoretical Background

There are different approaches to measuring and decomposing the dynamics of TFP. It can be measured by the index numbers such as the Divisia, Malmquist, Tornquist, Luenberger, and Fisher TFP indexes depending on the aggregation of outputs and inputs. The most commonly used measure is the Malmquist index, but a conventional measure is the Divisia index (Zhu et al. 2012). A method first proposed by Kumbhakar (1996) and Kumbhakar and Lovell (2003) decomposes TFP into technical change, scale effects, technical efficiency, and a price component. Following this approach, different papers decompose TFP change commonly using either Malmquist or Divisia Indices. For instance, Balk (2001) using the Malmquist index identifies four components of TFP change.

Technical change (TC) results from a shift in the cost frontier. TC captures the improvement in best practices through the adoption of new technologies. For instance, farmers using new crop varieties can produce more output at least cost. As a result, the best farms are getting better. TC can be positive or negative depending on whether the shift in the cost frontier down or up. The second component of TFP change is efficiency change (EC), the improvement in the firm's ability to use available technology. EC includes movement towards the cost frontier due to improved farm management, for example, or the wider adoption of better technology (Kumbhakar and Lovell 2003). The third component is the change in scale efficiency change (SC). SC shows movements along the cost function and a decrease in the average cost of production (Coelli et al. 2005). The fourth component is the input and the output mix effect (mixed-effect), which is very common in the multiple-input-multiple-output firm. The mix effect measures the effects of change in the composition of inputs and output vectors over time (Balk 2001).

Kumbhakar and Lozano-Vivas (2005) used the production frontier model to decompose the Divisia TFP growth into Technical efficiency change (TEC), technical change (TC), allocative efficiency change (AEC) and scale change (SC) components. On the other hand, Brümmer et al. (2002) decomposed the Divisa TFP change into TEC, TC, AEC and SC component using output distance function. Using input distance function Karagiannis et al. (2004) decompose Divisia TFP change into the same four components.

There have been several attempts to identify the relationship between profitability and productivity change. For instance, Miller and Rao (1989) decomposed profit

change into a productivity effect, an activity effect, and price effect. Grifell-Tatjé and Lovell (1999) developed an analytical framework in which profit change over time decomposed into price effect, an activity effect and productivity change effect. Activity effect includes resource mix, product mix, and scale effect. Productivity change effect includes operating efficiency and technical change effect. Kumbhakar and Lien (2009) decomposed the productivity effect further into technical efficiency and technological change effects while the activity effect subdivided into the scale, resource mix, and product mix effects see also (Sipilainen et al. 2013).

Our theoretical framework to a large extent follows the approach used by Kumbhakar and Lien (2009) and Sipilainen et al. (2013). In these studies, the dynamics of profitability change over time are measured as a change in profit based on the input distance function approach. These studies focused on the technical evaluation of dairy firms. However, the technical evaluation alone reveals how well farmers use the physical production process. There is a need to closely examine the cost efficiency of the farmers which will also address the management of financial resources. Moreover, Binswanger (1974) has shown that the dual approach is more desirable than the production function approach for economic analysis. The dual cost minimization framework is widely used in productivity literature to estimate and decompose productivity change through time (Kumbhakar and Lovell 2003). The theory of the cost function relies on the assumption that firms choose inputs to the production process that minimize the cost of producing output. The next subsection discusses measuring the level of productivity and profitability change using the dual approach.

2.2 Application

2.2.1 Productivity (TFP) Change Decomposition

Suppose we have a dataset of N firms over T periods and let $x_{it}=(x_{1it},\ldots,x_{nit})$ be the input quantity vector for firm i in period t and $X_{it} \equiv X(x_{it})$ be the aggregate input function. $y_{it}= (y_{1it},\ldots,y_{mit})$ is the output quantity vector for firm i in period t and $Y_{it} \equiv Y(y_{it})$ is the aggregate output function. where X and Y are non-negative, non-decreasing and linearly homogenous aggregator functions. Output quantities are measures of quantities sold plus on-farm consumption and net changes in inventories. Input quantities are measures of purchasing inputs as well as farm production used on the farm. If a technology produces multiple outputs, TFP change ($T\dot{F}P$) is defined as the difference between the rate of change of an output index ($\dot{Y}$) and the rate of change of an input index ($\dot{X}$) (Kumbhakar et al. 2014). For the development of expressions (1) to (6), we will suppress the firm subscript i.

$$T\dot{F}P = \dot{\boldsymbol{Y}} - \dot{\boldsymbol{X}} \equiv \sum_m R_m \dot{y}_m - \sum_j S_j \dot{x}_j \tag{1}$$

where a dot above a variable will denote the rate of change in the log of that variable; $R_m = p_m y_m/R$, $R=\sum_m P_m y_m$ in which R is total revenue and R_m is the observed revenue share of output y_m; p is the output price vector ($p=p_1,\ldots,p_m$); y is the vector of output; and S_j is the observed expenditure share of input X_j ($S_j = w_j x_j/C$). C is the total cost ($C = \sum_j w_j x_j$); and w is the vector of input price ($w=w_1,\ldots,w_j$). As shown by Kumbhakar and Lien (2009) and Sipilåinen et al. (2013) Eq. (1) can be re-written as:

$$\dot{TFP} = \mathrm{TC} + \mathrm{EC} + \left(1 - RTS^{-1}\right)\dot{y}_c + \left(\dot{y}_p - \dot{y}_c\right) \equiv TC + EC + Scale + Markup \tag{2}$$

where $\mathrm{TC} = -\frac{\partial lnC}{\partial t}$; $RTS^{-1} = \sum_m \frac{\partial lnC}{\partial \ln y_m}$, $\dot{y}_c = RTS\left[\sum_m \frac{\partial lnC}{\partial \ln y_m}\dot{y}_m\right]$, $\dot{y}_p = \sum_m R_m \dot{y}_m$, and $\dot{y}_m$ is the rate of change in output y_m. EC (efficiency change) $= \frac{\partial TE}{\partial t}$; TE is the mean efficiency level of the firm at a given time. RTS is returns to scale of the firm. Using this concept we can decompose the profitability change in the next subsection.

2.2.2 Profitability Change Decomposition

A profit of a firm (π) = *Revenue* (R) − *cost* (C) $= \sum_m p_m y_m - \sum_j w_j x_j$ and change in profit using Eq. 2 is expressed as:

$$\frac{d\pi}{dt} = \left(\sum_m p_m \frac{\partial y_m}{\partial t} + \sum_m y_m \frac{\partial p_m}{\partial t}\right) - \left(\sum_j w_j \frac{\partial x_j}{\partial t} + \sum_j x_j \frac{\partial w_j}{\partial t}\right)$$

$$= \left(\sum_m p_m y_m \frac{\partial lny_m}{\partial t} + \sum_m y_m p_m \frac{\partial lnP_m}{\partial t}\right) - \left(\sum_j w_j x_j \frac{\partial lnx_j}{\partial t} + \sum_j x_j w_j \frac{\partial lnw_j}{\partial t}\right)$$

$$= R\left(\sum_m R_m \dot{y}_m + \sum_m R_m \dot{p}_m\right) - C\left(\sum_j S_j \dot{x}_j + \sum_j S_j \dot{w}_j\right) \tag{3}$$

Divide Eq. (3) by total cost

$$\frac{1}{C}\frac{d\pi}{dt} = \frac{R}{C}\left(\sum_m R_m \dot{y}_m + \sum_m R_m \dot{P}_m\right) - \left(\sum_j S_j \dot{x}_j + \sum_j S_j \dot{w}_j\right) \tag{4}$$

From Eq. (1) and (2) we can get

$$-\sum_j S_j \dot{x}_j = TC + EC + \left(1 - RTS^{-1}\right)\dot{y}_c + \left(\dot{y}_p - \dot{y}_c\right) - \sum_m R_m \dot{y}_m \tag{5}$$

Substituting (5) into (4)

$$\frac{1}{C}\frac{d\pi}{dt} = \frac{R}{C}\left[\sum_m R_m\dot{y}_m + \sum_m R_m\dot{p}_m\right] - \sum_j S_j\dot{w}_j + \quad TC + EC + \left(1 - \sum_m \frac{\partial lnC}{\partial \ln y_m}\right)\dot{y}_c + \left(\dot{y}_p - \dot{y}_c\right) - \sum_m R_m\dot{y}_m$$

$$\frac{1}{C}\frac{d\pi}{dt} = \left(\frac{R}{C} - 1\right)\dot{y}_p + \left(\frac{R}{C}\right)\dot{p} - \dot{w} + TC + EC + \left(1 - RTS^{-1}\right)\dot{y}_c + \left(\dot{y}_p - \dot{y}_c\right) \tag{6}$$

$$\frac{1}{C}\frac{d\pi}{dt} \equiv \left(\frac{R}{C} - 1\right)\dot{y}_p + \left(\frac{R}{C}\right)\dot{p} - \dot{w} + T\dot{F}P$$

where $\dot{p} = \sum_m R_m\dot{p}_m$ *and* $\dot{w} = \sum_j S_j\dot{w}_j$. Equation (6) is of primary interest for this study, which decomposes the change in profit as a percentage of total cost into several components. Following Kumbhakar et al. (2009) and Sipilainen et al. (2013), we can give an interpretation of each component in (6) as follows:

(a) TC is the technical change component $\left(-\frac{\partial lnC}{\partial t}\right)$, which will affect profitability positively if there is technical progress;
(b) $\left(1 - RTS^{-1}\right)\dot{y}_c$ is the scale component and measures the effect scale economies. It will increase profit if RTS >1 and the aggregate output cost ($\dot{y}_c$) is small.
(c) $\dot{y}_p - \dot{y}_c$ is the markup component. It will increase profitability of the farm if the markup change is positive.
(d) EC is the efficiency change component (EC $= \frac{\partial u}{\partial t}$), which will affect profit positively if efficiency improves over time;
(e) $\dot{y}_p\left(\frac{R}{C} - 1\right)$ is the output growth component, which will increase profitability if the output growth rate is positive.
(f) $\left(\frac{R}{C}\right)\dot{p}$ is the output price change component; which will affect profit positively if output price increase overtime;
(g) $\dot{w}$ is input price change component; which will affect profit positively if input price change is negative

Output change, input and output price change components can be computed simply from the data, while TC, scale, markup, and EC require econometric estimation.

3 The Econometric Model

A cost function gives the minimum[2] cost of producing a given level of output given input prices and technology. That is, we assume that a firm i (i=1,...,N) is a cost-minimizing entity that produces output Y subject to a production constraint $F = (\boldsymbol{Y}, \boldsymbol{X})$. The mathematical expression as follows:

$$\text{Min } C = \sum_{j=1}^{n} W_j X_j \left(\boldsymbol{Y}, \boldsymbol{W}\right) \tag{7}$$

Subject to

$$F\left(\boldsymbol{Y}, \boldsymbol{X}\right) = 0$$

The true cost function is unknown. Thus, consistent with most of the firm efficiency literature (Christensen and Greene 1976), we can estimate a Transcendental Logarithmic (TL) cost function. It is continuous and non-negative, as well as positively linearly homogenous, non-decreasing, and concave on price; non-decreasing, and quasi-convex on output. Our specification of a multi-product TL cost function C for j=1,..., J inputs and m=1,..., M outputs can be specified in log form as:

$$\begin{aligned} lnc = \alpha_0 + \sum_{j=2}^{4} \beta_j \ln \check{w}_j + \sum_{m=1}^{3} \propto_m \ln y_m + \sum_{l=1}^{3} \propto_{mt} \ln y_m t + \sum_{j=2}^{4} \beta_{jt} \ln \check{w}_j t + \\ + \frac{1}{2}\left[\sum_{m=1}^{3}\sum_{m=1}^{3} \gamma_{mm} \ln y_m \ln y_m + \sum_{j=2}^{4}\sum_{j=2}^{4} \delta_{jj} \ln \check{w}_l \ln \check{w}_j + \rho t^2\right] + \\ \sum_{j=2}^{3}\sum_{m=1}^{3} \varnothing_{jm} \ln y_m \ln \check{w}_j + dt + \mu_i + V_{it} + U_{it} \end{aligned} \tag{8}$$

where lnc represents log form of total cost, w_j represent the price of inputs j, and y_i is the quantity of output i. $\ln \check{w}_j = \ln w_j - \ln w_1 \ (\forall j)$ discussed in the next paragraph. All Greek letters are parameters to be estimated and the white noise error term (V_{it}) is added to allow for random measurement error in Eq. (8). μ_i capture latent heterogeneity (farm-effect). U_{it} is the non-negative variable representing technical inefficiency. We assumed V_{it} is symmetric and to satisfy the classical assumptions, i.e. ${v_{it}}^{iid} \sim N\left(0, \sigma_v^2\right)$, $V_{it} \perp U_{it}$. The trend variable, t, include

[2]In a cost minimization setup the output(y) is treated as exogenous and the inputs (x) are treated as endogenous.

to capture Hicks-neutral technology change starts with T=91 for 1991 and increases by one annually. Economic theory imposes homogenous and symmetry restrictions on the parameters. Any sensible cost function must be homogenous of degree 1 in input prices; thus the restrictions in input prices $\sum_j^k \beta_j = 1, \sum_j^k \gamma_{jl} = \sum_j^k \delta_{jl} = 0$;and the symmetry restriction $\gamma_{lj} = \gamma_{ji}$. From Eq. (8) we can derive the cost share function (S_j) using Shephard's lemma as follows:

$$s_j = \frac{\partial lnc}{\partial \ln w_j} = \frac{w_j x_j}{c} = \beta_j + \sum_{j=2}^{l} \delta_{jj} \ln \check{w}_l + \sum\nolimits_{l=1}^{m} \varnothing_{jm} \ln y_m + \beta_{jt} t \qquad (9)$$

Since $\sum_{j=1}^{j} s_j = 1$, the cost share Eq. (9) must satisfy the adding-up property. However, this property implies the same restrictions as linear homogeneity in the cost function, so we imposed both properties by dividing the quantity of all inputs by the quantity of one of the inputs. Then, in Eq.(9) we imposed homogenous restriction by re-defining both the left- and right-hand sides of the equations as follows: $\ln \check{w}_j = \ln w_j - \ln w_1 \ (\forall j)$ and $lnc = \ln\left(c / w_1\right)$. This approach also implies that one of the share equations has to be dropped. The parameters of the dropped equation can be recovered from the homogeneity restrictions discussed above. Using Eqs. (8) and (9), we computed the seven components of profitability change shown in Eq. (6). We used Greene (2005) model to estimate parameters in Eq. (8). The next section discusses data source and variables.

4 Data and Definition of Variables

The data used in this analysis is an unbalanced panel with 3885 observations from farmers involved only in the production of crops (grain and forage) for the year 1991–2013. The data include production and economic data collected annually by the Norwegian Institute of Bioeconomy Research (NIBIO) from about[3] 1000 farms in all regions of Norway. Participation in the survey is voluntary. There is no limit on the number of years a farm included in the study. Some of the farmers participated more than 20 years, and others have started participating for the first time. To accommodate panel features in estimation, we included only those farms for which at least three consecutive years of data are available.

The output measure at our disposal in the data set is the grain output in1000 FU[4] (y_1), forage output in1000 FU (y_2), and other crop outputs in 1000 in Norwegian

[3]The number of participants varies from year to year. For example in 1991 data has been collected from 1049 firms but in 2013 it was 924 firms. Approximately 10% of the survey farms are replaced per year to incorporate changes in the population of farms in Norway.

[4]FU stands for feed units, which adjust the quality difference in output. 1 FU = 1 kg of grain with the 15% water content. Thus the output is quality-adjusted yield in kilograms per year.

Kroner (NOK) (y_3). Grain output is an aggregate of four main species: barley, wheat, oats, and oilseed species. The aggregate is quality adjusted and is measured in FU (feed units) as defined by NIBIO. Thus, the natural output to use is the quality-adjusted crop output in kilograms per decare (daa).[5]

To assess the efficiency and productivity growth, we need to be sure that farmers under consideration are comparable. Forage and grain output can be an input for livestock production so that it can be an intermediate product. To avoid double counting, we have selected only 455 farmers who are involved producing grain, forage and other crop products (potatoes, tomatoes, vegetables, etc.). These firms are located in the eastern and central (Trøndlag) regions of Norway. Moreover, we exclude government intervention like a subsidy in the main output because the main task of the research is to know how the farmers allocate resources to produce crop production. Several studies conducted on the effect of subsidizing conclude that government farm support distorts efficiency (for instance Kleinhanß et al. 2007; Kumbhakar and Lien, 2009). Output prices (P_m) corresponding to the output variables are estimated from the survey data. Implicit output prices are calculated from output revenue for each kind of crop divided by the output quantity for each crop type. Prices for other outputs are aggregated as a Fisher index (Diewert 1998).

Major inputs include labor, measured as the total labor hours used in the farm, including hired labor, owners' labor, and family labor; farmland, defined as productive land (both owned and rented); material which includes inputs such as fertilizer, seed, and pesticide, registered by their costs of purchase in NOK; and capital is measured as the sum maintenance and running (hiring) costs, depreciation and interest costs on the total capital stock (3%) deflated by an index for fixed cost items figure from NIBIO and calculated at 2013 price levels.

The cost function (8) is specified with the following four input prices (w_j). Land prices are derived from the market prices for rental of farmland, in the area of each farm. The price of labor is the wage of hiring labor. The price of other variable inputs and capital costs were constructed as Laspeyres indices based on figures provided by NIBIO. All prices are deflated to 2013 levels using the agricultural price index figures also provided by NIBIO. Descriptive statistics of data are summarized in Table 1. Norwegian farmers are small. The annual average output was about 61,000 FU of grain and 77,000 FU of forage. The average farm received on the output grain price of 1.86 Norwegian kroner (NOK) per FU and 0.33 NOK per FU for forage. Figure 1 shows crop output per year was increasing in all three agricultural outputs and follows an almost similar trend.

[5] A decare (daa) is equal to 0.1 hectare (ha).

Table 1 Descriptive statics of model variables in cost function

Variables	Label	Unit	Mean	Std. Dev.
Output (y_i) y_1	Grain output	1000 FU	60.829	65.715
y_2	Forage output	1000 FU	76.657	54.278
y_3	Other	1000 NOK	9.305	7.546
Inputs (x_i) x_1	Labor	1000 h	3.492	1.238
x_2	Land	1000 daa	0.346	0.202
x_3	Material cost	1000 NOK	217.639	133.380
x_4	Capital cost	1000 NOK	352.743	343.190
Inputs (p_i) p_1	Grain price	NOK/FU	1.855	0.559
p_2	Forage price	NOK/FU	0.324	0.477
p_3	Other crop prices	index	62.836	11.921
Inputs price (wj) w_1	Wage	NOK/h	144.496	31.926
w_2	Rent	NOK/daa	237.053	154.551
w_3	material price	index	67.187	15.590
w_4	capital price	index	80.741	10.359
T	Trend (1=year 1991)			
N	Sample size		3885	

NOK Norwegian Kroner and *FU* Feed Units

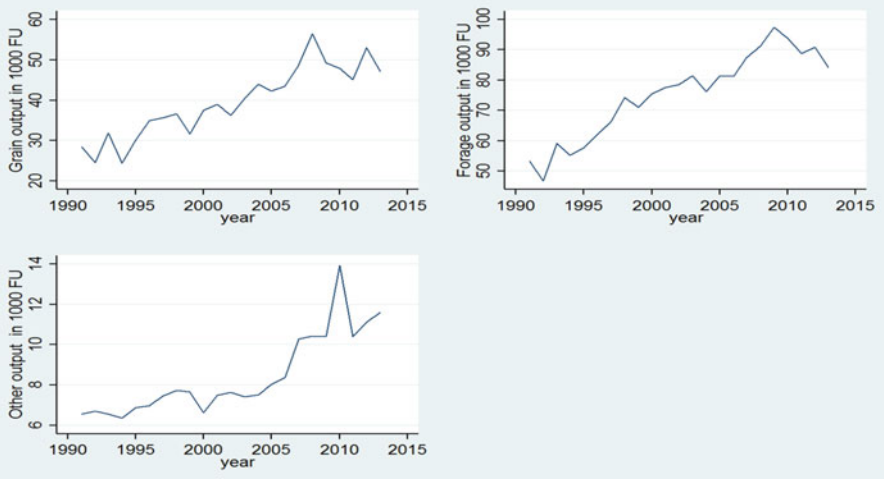


Fig. 1 Annual mean crop output from 1991–2013

5 Estimation and Results

5.1 Testing Model Specification

The cost function is estimated using STATA® version14. The trend variable is normalized to be zero in the year 2013. All other variables normalized before taking the logarithms by dividing each variable by its mean value so that the first-order parameters can be interpreted as elasticities at geometric mean. The estimated parameters and associated standard errors are reported in Table 3. The results show that the estimated variable cost function is not decreasing on each input price and output quantity at any reasonable level of significance. Various specification tests were conducted to obtain the best model and functional form for the data under analysis (Table 2).

Before estimating the production function, the skewness of the data tested based on Schmidt & Lin (1984). The test return of skewness with a P value less than 0.001 shows that the null hypothesis of no skewness confidently rejected. The null hypothesis that there are no technical efficiency effect in the models was tested. The null hypothesis rejected, in which the LR is greater than the (mixed) chi-square value of 5.412. A generalized likelihood ratio test using a mixed chi-squared distribution is consistent technical inefficiency constituting the largest share of total error variance, suggesting the appropriateness of the stochastic frontier analysis (SF) approach as opposed to ordinary least squares (OLS). Moreover, likelihood function expressed in terms of the two variance parameters as $\gamma = \sigma_u^2/\sigma_u^2 + \sigma_v^2$ ($\gamma = 0.34$ in Table 3) shows that technical inefficiency consist the largest share of total error variance supports the appropriateness using SF approach. An LR tests reject a simplification of the TL to CD rejected. The goodness of fit measured by the log of likelihood function is statistically significant.

Table 2 Properties of grain and forage production technology

Restrictions	Parametric restrictions	chi2	*p-value*
Cobb-Douglas technology	H_0: All interaction terms are zero	1285	0.000
Scale technology effects in output	H_0: $\propto_{mt} = 0$	12.31	0.006
Hicks technology effects in inputs	H_0: $\beta_{jt} = 0$	50.63	0.000
Schmidt & Lin (1984)	H_0: no skewness	445.8	0.000
Generalized LR ratio test	Test for one-sided error	61.02 [a]	0000

[a]Denotes significant at 1% level of significance using mixed chi-square distribution with 1 degree of freedom and a critical value of 5.412

Table 3 Estimates of the parameters of the translog stochastic frontier model

Variable	First orders	$\ln\check{w}_2$	$\ln\check{w}_3$	$\ln\check{w}_4$	$\ln y_1$	$\ln y_2$	$\ln y_3$	t
Constant	−0.21***(0.02)							
$\ln\check{w}_2$	0.04***(0.01)	0.02*(0.01)						
$\ln\check{w}_3$	0.35*** (0.14)	0.01(0.11)	9.70**(3.71)					
$\ln\check{w}_4$	0.54***(0.14)	−0.07(0.12)	0.07(0.05)	−15.95*(7.95)				
$\ln y_1$	0.14***(0.01)	0.01(0.01)	0.05(0.01)	0.02*(0.01)	0.04***(0.01)			
$\ln y_2$	0.13***(0.01)	0.01(0.01)	−0.18*(0.09)	0.19*(0.09)	0.01(0.01)	0.03***(0.01)		
$\ln y_3$	0.23***(0.01)	−0.01(0.01)	0.13(0.09)	−0.12(0.10)	−0.03***(0.01)	0.002(0.00)	0.05***(0.01)	
t	0.03***(0.00)	0.00*(0.00)	−0.32*(0.13)	0.31*(0.13)	0.00***(0.00)	0.00***(0.00)	0.00***(0.00)	0.008***(0.00)
Log. L.= 1040***	$\gamma = \frac{\sigma_u^2}{\sigma_u^2+\sigma_v^2} = 0.34$		N=3885	(0.00)				

Note: Standard errors in parentheses, * $p<0.05$, ** $p<0.01$, *** $p<0.001$; and $\check{w}_2$=land/labour, $\check{w}_3$=var.cost/labor,$\check{w}_4$=fixed.cost/labour, y_1= grain, y_2=forage, and y_3=other, all in log form

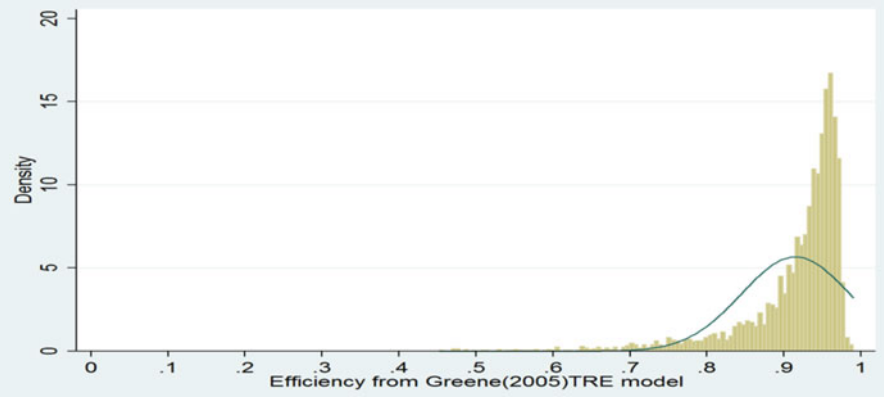


Fig. 2 Histogram of the efficiency index from the Greene (2005) model (The solid line is the fitted value for the model)

5.2 Inefficiency Score

We plot a histogram of efficiency index using the Greene (2005) true random-effects model (TRE) model (Fig. 2). The estimated efficiency score across the years of observation is 0.91. The estimated efficiency index implies that the minimum cost is about 91% of the actual expenditure. Alternatively, the actual cost can be reduced without reducing the output by 10% (1/0.91–1) if we remove inefficiency in crop production in Norway.

5.3 Price and Output Elasticities

Table 3 shows the parameters of stochastic frontier model estimation. The models exhibit positive and highly significant first-order parameters, fulfilling the monotonicity condition as expected for a well-behaved cost function. The elasticity of cost on the price of land, other variable input costs and capital costs were 0.04, 0.35 and 0.54, respectively. If one percent increases the price of land, costs will increase by an estimated 0.04%, ceteris paribus. If the price of other variable inputs increases by 1%, costs will increase by an estimated 0.35%. The coefficient for the capital (fixed input) price (0.54) is the largest among other partial elasticities and statistically significant ($p < 0.001$). The result implies that crop production in Norway more of capital intensive and the percentage change in the capital price has a larger influence on crop production compared to other inputs. Thus, any intervention to improve the crop sector needs to prioritize on these inputs. We can recover and estimate the elasticity of cost on the price of labor, i.e., If the price of labor increases by 1%, costs will increase by an estimated 0.07, i.e. $1-(0.04+0.35+0.54)$. The elasticity

of cost on grain, forage, and other outputs were 0.14, 0.13 and 0.23, respectively. This means for instance, if grain output increases by 1%, costs increased by an estimated 0.14%, ceteris paribus.

5.4 TFP and Profitability Change

The components of TFP and profitability change are plotted in Figs. 3 and 4, respectively. The estimated average TFP and profitability change are reported in Table 4. The result indicates that the overall average annual change in the TFP growth rate in grain and forage production during the period 1991–2013 was −0.11% per annum. This result is consistent with the results from previous studies. For instance, a survey conducted for Polish Agriculture reported TFP decreased by 2% over the period 1996 to 2000 (Latruffe et al. 2008) Moreover, Baráth and Fertő (2017) reported a decline in TFP for European agriculture from 2004 to 2013. TFP decline was mainly due to negative contributions from the markup change component. There are no similar studies conducted for multiple output technology in forage and grain production for comparison. The estimated result shows that technological change (TC) was – 0.03% per annum. Moreover, Wang and Ho (2010) stated that the first order coefficients of the time trend variable show estimates of the

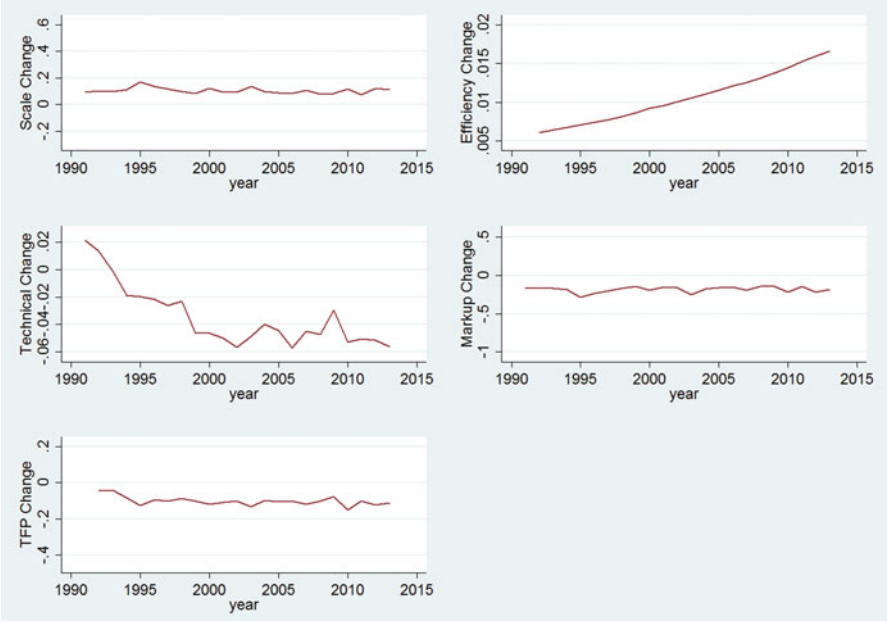


Fig. 3 Mean TFP change components estimated from cost function for the year 1991–2013

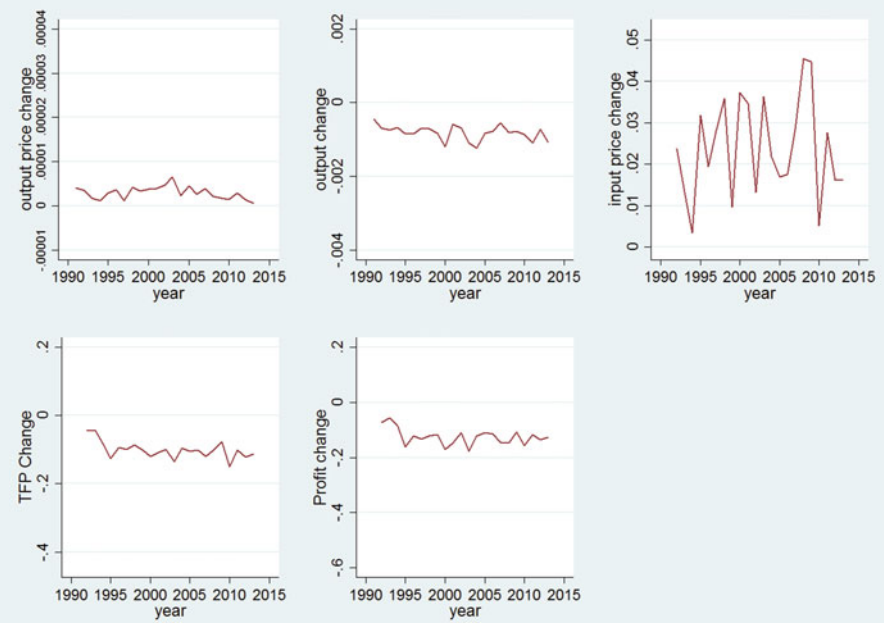


Fig. 4 Mean profit change components for the year 1991–2013

Table 4 Annual TFP, profitability change and its components (in percent)

Variable	Mean	Std. Dev.
Efficiency change (EC)	0.0102	0.0031
Scale change(SC)	0.1001	0.4775
Technical change (TC)	−0.0327	0.0228
Markup	−0.1859	0.7348
TFP change	−0.1122	0.2829
Output price change	0.00002	0.00005
Input price change	0.0259	0.0195
Output change	−0.0011	0.0035
Profit change	−0.1379	0.2867
Sample size (N)	3885	

average annual rate of technical change (TC). The estimated parameter of the trend variable is positive and statically different from zero at the 1% level of significance, which suggests technical regress for Norwegian crop production during the study period. A similar result reported in other studies, for instance, in the Latruffe et al. (2008) study cited above for Poland for the period 1996–2000. The quadratic term in Table 3 is positive (coefficient of t^2), indicating that the technical regress is getting stronger over time.

Technical regress may be explained by several factors as discussed in Kumbhakar and Heshmati (1995) and Kumbhaker et al. (2008). First, they argue that changes in the regulations concerning input use, for instance, government controls on the use

of pesticides and fertilizers. Second, there might be increased technical inefficiency over time due to lack of external competition and restrictions on the transfer of farms between generations. Finally, a larger real increase in input prices than output prices may lead to results that look like technical regress. In another study Atsbeha et al. (2015) proposed that such apparent technical regress can occur if the sector under consideration is subject to scale restrictions, results in the scale of production that is sub-optimal to the latest technologies. The later reason is relevant for the Norwegian crop producing farms, in which the sector is based on small scale family farms and the land is fragmented. Thus, if there are economies of scale in the production of crop-producing technologies, this development may result in a shift in the long-run average cost of crop producing farms that leave small farms worse off over time.

As shown in Table 4 in the appendix, technical regress has been neither scale nor Hicks neutral. Moreover, Hicks non-neutrality of technology regress is exhibited as a significant interaction parameter with time (t) for cost share of variable inputs ($\check{w}_3$) and fixed input ($\check{w}_4$) (Table 2). Technological change exhibits a positive effect on the cost share of fixed inputs and a negative effect for cost share of variable inputs. Thus TC was non-neutral over the last 23 years. With respect to scale, the interaction parameter with time (t) for grain production ($\text{tln}y_1$) is negative and statically significant, which suggests that the cost increasing effects of technical regress get weaker as grain production increase. However, the interaction parameter with time (t) for other output ($\text{tln}y_3$) is positive and statistically significant, which suggests that the cost increasing effects of technical regress have become stronger for the other output production increase. These suggest that the technical regress was more important for small scale grain production and big scale other output production.

Efficiency change, which measures the change from observed cost towards the best practice farmers, was positive (0.01) % per annum. The estimated result of a decline in TC and an improvement in efficiency shows farmers are able to adopt the prevailing technology and hence lie, on average, closer to the frontier (Latruffe et al. 2008). The scale component (SC) was positively contributed to the total productivity change (0.10) % per annum. The contribution of the markup for the period 1991–2013 was −0.19% per annum. A markup effect could show firms have some market power and price above their marginal cost. However, a negative markup implies that market power through price-making does not give effect on firms' performance. A non-zero markup effect on TFP means that output prices diverge from the marginal cost of production, i.e. the output market is non-competitive (Sipiläinen et al. 2013). The contribution markup effects are shown in Fig. 1, which is fluctuating considerably over time and has almost the same movement as that of TFP change.

Figure 4 and Table 4 shows the estimated results of annual profitability change component (in percent). The profit change component was −0.14, suggesting that profit has declined by 0.14% per annum. This is mainly because of the negative TFP change of 0.11% per year with some contributions from an input price annual change of 0.03%. The contributions from output change and output price change, which might have a positive effect on the profitability change, are almost zero.

6 Discussion and Conclusion

The economic performance of a farm is commonly measured by the efficiency and productivity measures. We used farm level unbalanced panel data for the year 1991–2013. We have selected only crop producing specialized 455 farms located in the eastern and central (Trøndlag) regions of Norway. We have estimated the profitability and productivity of the Norwegian crop producing specialized farms using a translog cost function. The result indicates that average annual TFP growth rate in grain and forage production declined by 0.11% per annum during the period 1991–2013. The contributions of the technological change and markup change were − 0.03% and −0.19% per annum, respectively. Efficiency change, which measures the change from observed cost towards the best practice farmers, was positive (0.01% per annum). Moreover, the scale component has positively contributed to the total productivity change (0.04% per annum). The profit change declining by 0.14% and this was mainly because of the negative TFP change with some contributions from an input price change component increase by 0.03% per year.

Technical change captures the shift in technology and is the key driver of profitability and productivity growth. Policy makers have to give priority to investing in agricultural research and development, which can help in the innovation of new technologies and improvement in TC. Investment in research and development support for innovation of new technology and improves TC (O'Donnell 2010). The study also shows that there was a small efficiency change for the last 23 years. Thus, farmers continue to be lagging behind the best-practice farmers. Therefore, there is a need for intensive work on agricultural extension and dissemination to help farmers adopt the existing technologies.

Acknowledgements The Norwegian research council, grant number 225330/E40, supported the project. I am grateful for the financial assistance of the Research Council of Norway. I thank the reviewer for his/her thorough review and valuable comments, which significantly contributed to improving the quality of the paper.

References

Atsbeha, D. M., Kristofersson, D., & Rickertsen, K. (2015). Broad breeding goals and production costs in dairy farming. *Journal of Productivity Analysis, 43*(3), 403–415.

Balk, B. M. (2001). Scale efficiency and productivity change. *Journal of Productivity Analysis, 15*, 159–183.

Baráth, L., & Fertő, I. (2017). Productivity and convergence in European agriculture. *Journal of Agricultural Economics, 68*(1), 228–248.

Binswanger, H. P. (1974). A cost function approach to the measurement of elasticities of factor demand and elasticities of substitution. *American Journal of Agricultural Economics, 56*(2), 377–386.

Brümmer, B., Glauben, T., & Thijssen, G. (2002). Decomposition of productivity growth using distance functions: the case of dairy farms in three European countries. *American Journal of Agricultural Economics, 84*(3), 628–644.

Christensen, L. R., & Greene, W. H. (1976). Economies of scale in US electric power generation. *The Journal of Political Economy, 84*(4), 655–676.

Coelli, T. J., Rao, D. S. P., O'Donnell, C. J., & Battese, G. E. (2005). *An introduction to efficiency and productivity analysis*. New York: Springer Science & Business Media.

Diewert, W. E. (1998). Index number issues in the consumer price index. *The Journal of Economic Perspectives, 12*(1), 47–58.

Greene, W. (2005). Fixed and random effects in stochastic frontier models. *Journal of Productivity Analysis, 23*(1), 7–32.

Grifell-Tatjé, E., & Lovell, C. K. (1999). Profits and productivity. *Management Science, 45*(9), 1177–1193.

Karagiannis, G., Midmore, P., & Tzouvelekas, V. (2004). Parametric decomposition of output growth using a stochastic input distance function. *American Journal of Agricultural Economics, 86*(4), 1044–1057.

Kleinhanß, W., Murillo, C., San Juan, C., & Sperlich, S. (2007). Efficiency, subsidies, and environmental adaptation of animal farming under CAP. *Agricultural Economics, 36*(1), 49–65.

Koesling, M., Flaten, O., & Lien, G. (2008). Factors influencing the conversion to organic farming in Norway. *International Journal of Agricultural Resources, Governance and Ecology, 7*(1), 78–95.

Kumbhakar, S. C. (1996). Efficiency measurement with multiple outputs and multiple inputs. *Journal of Productivity Analysis, 7*(2), 225–255.

Kumbhakar, S. C., & Heshmati, A. (1995). Efficiency measurement in Swedish dairy farms: an application of rotating panel data, 1976–88. *American Journal of Agricultural Economics, 77*(3), 660–674.

Kumbhakar, S. C., & Lien, G. (2009). Productivity and profitability decomposition: A parametric distance function approach. *Food Economics – Acta Agriculturae Scandinavica, Section C, 6*(3–4), 143–155.

Kumbhakar, S. C., & Lovell, C. K. (2003). *Stochastic frontier analysis*. Cambridge: Cambridge University Press.

Kumbhakar, S., & Lozano-Vivas, A. (2005). Deregulation and productivity: The case of Spanish banks. *Journal of Regulatory Economics, 27*(3), 331–351.

Kumbhakar, S. C., Lien, G., Flaten, O., & Tveterås, R. (2008). Impacts of Norwegian milk quotas on output growth: A modified distance function approach. *Journal of Agricultural Economics, 59*(2), 350–369.

Kumbhakar, S., Lien, G., & Hardaker, J. B. (2012). Technical efficiency in competing panel data models: a study of Norwegian grain farming. *Journal of Productivity Analysis, 41*(2), 321–337. https://doi.org/10.1007/s11123-012-0303-1.

Kumbhakar, S., Wang, H.-J., & Horncastle, A. (2014). *A practitioner's guide to stochastic frontier analysis using stata*. Cambridge: Cambridge University Press.

Latruffe, L., Davidova, S., & Balcombe, K. (2008). Productivity change in Polish agriculture: An illustration of a bootstrapping procedure applied to Malmquist indices. *Post-Communist Economies, 20*(4), 449–460.

Lien, G., Kumbhakar, S. C., & Hardaker, J. B. (2010). Determinants of off-farm work and its effects on farm performance: the case of Norwegian grain farmers. *Agricultural Economics, 41*(6), 577–586.

Melfou, K., Theocharopoulos, A., & Papanagiotou, E. (2007). Total factor productivity and sustainable agricultural development. *Economics and Rural Development, 3*(1), 32–38.

Miller, D. M., & Rao, P. M. (1989). Analysis of profit-linked total-factor productivity measurement models at the firm level. *Management Science, 35*(6), 757–767.

O'Donnell, C. J. (2010). Measuring and decomposing agricultural productivity and profitability change. *Australian Journal of Agricultural and Resource Economics, 54*(4), 527–560.

Odeck, J. (2007). Measuring technical efficiency and productivity growth: A comparison of SFA and DEA on Norwegian grain production data. *Applied Economics, 39*(20), 2617–2630.

Schmidt, P., & Lin, T.-F. (1984). Simple tests of alternative specifications in stochastic frontier models. *Journal of Econometrics, 24*(3), 349–361.

Sipilainen, T., Kumbhakar, S. C., & Lien, G. (2013). The performance of dairy farms in Finland and Norway from 1991 to 2008. *European Review of Agricultural Economics, 41*(1), 63–86.

Statistics-Norway. (2016). Agriculture, forestry, hunting, and fishing. Retrieved from Jan 25 2017. https://www.ssb.no/en/jord-skog-jakt-og-fiskeri

Wang, H.-J., & Ho, C.-W. (2010). Estimating fixed-effect panel stochastic frontier models by a model transformation. *Journal of Econometrics, 157*(2), 286–296.

Zhu, X., Demeter, R. M., & Lansink, A. O. (2012). Technical efficiency and productivity differentials of dairy farms in three EU countries: The role of CAP subsidies. *Agricultural Economics Review, 13*(1), 66.

Estimation of Health Care Demand and its Implication on Income Effects of Individuals

Hossein Kavand and Marcel Voia

Abstract Zero inflation and over-dispersion issues can significantly affect the predicted probabilities as well as lead to unreliable estimations in count data models. This paper investigates whether considering this issue for German Socioeconomic Panel (1984–1995), used by Riphahn et al. (2003), provides any evidence of misspecification in their estimated models for adverse selection and moral hazard effects in health demand market. The paper has the following contributions: first, it shows that estimated parameters for adverse selection and moral hazard effects are sensitive to the model choice; second, the random effects panel data as well as standard pooled data models do not provide reliable estimates for health care demand (doctor visits); third, it shows that by appropriately accounting for zero inflation and over-dispersion there is no evidence of adverse selection behaviour and that moral hazard plays a positive and significant role for visiting more doctors. These results are robust for both males and females' subsamples as well as for the full data sample.

Keywords Over-dispersion · Zero-inflated distribution · Adverse selection · Moral hazard · Health demand

JEL Classification: C46, C52, I11, I13

1 Introduction and Literature Review

Pauly (1968), Rothschild and Stiglitz (1976), and Bundorf et al. (2005), respectively, delineate the effect of moral hazard, adverse selection, and income effect in health insurance markets. A number of studies have investigated these effects, though

H. Kavand • M. Voia (✉)
Carleton University, Ottawa, ON, Canada
e-mail: hosseinkavand@cmail.carleton.ca; marcel.voia@carleton.ca

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_13

sometimes different econometric methodologies led to different interpretations about the effects of similar data.

Following the model developed by Cameron et al. (1988), Riphahn et al. (2003) we estimate the demand for doctor and hospital visits for the German Socioeconomic Panel data (GSOEP, 1984–1995). The findings of this study suggest that adverse selection, where a high-risk individual buys more insurance coverage, affects positively the number of doctor and hospital visits for only the males' hospital demand. Moral hazard, where an insured individual uses more health care services because of its lower cost, does not have any significant effect on any of the above health-care demands. Among other studies that looked at the effect of asymmetry information on health care demand are the studies of Chiappori and Salani (2000) that examined adverse selection using German data, while Geil et al. (1997) and Cameron et al. (1988) investigate moral hazard using Australian data.

In their theoretical model, Wolfe and Goddeeris (1991) delineate that wealth can ambiguously affect health-care demand. Their empirical results, however, indicate that both "bequeathable" and "non-bequeathable" wealth substantially increase the demand for both supplementary health care and health expenditure. Data reveals that those who enrolled for supplementary insurance, on average, had 50% higher wealth. They estimate the effect of moral hazard in a health expenditure model and use its estimated error term as a proxy for unexplained expenditure in their health care demand model. Its significant coefficient indicate the existence of selection effect.

For U.S. data, Marvasti (2014) finds that the demand for services of doctors is neither income elastic nor price elastic. As Marvasti discusses, however, Bago d'Uva and Jones' (2009) latent hurdle model confirms a positive income effect on the number of visiting doctors for European data. Amponsah (2013) confirms moral hazard and adverse selection for Ghanaian health care, and finds that an individual in a higher income group to be more likely to buy health insurance (income effect). His study confirms the income effect found in Asante and Aikins (2008) and Kirigia et al. (2005).

Keane and Stavrunova (2016) use a simultaneous equation model to jointly investigate adverse selection and moral hazard for the U.S. supplemental health-insurance market, namely Medigap. They extend previous studies by employing a smooth mixture of the Tobit model to control for heterogeneity, and by capturing the correlation between unobservable factors that affect both health insurance demand and health expenditure. Although they find a negligible adverse selection into Medigap, the insurance coverage leads to a significant rise in health care utilization and its related costs (moral hazard). Conditional on the supplemental insurance and health status, income has a small effect on health expenditure (See Cardon and Handel (2001) and Bajari et al. (2011, 2014) for other health market structural models).

To deal with inflated zeros in health care demand variables, researchers rely on different approaches. Powell and Goldman (2016) control for zero medical care expenditure and its heavily-skewed distribution by employing a quantile treatment effect framework (see Powell (2014)). The framework estimates an unconditional distribution for health care expenditures by assuming no adverse selection. They compare this distribution with the observed health expenditure distribution to estimate adverse selection effects. After separating moral hazard and adverse selection, each factor almost explains half of the reason for higher medical expenditure of a most generous plan compared to a least one.

With around 90% and 35% zeros for the number of doctor visits and hospital visits, the results of Riphahn et al. (2003) can also suffer from over-dispersion by not taking them into account. To mitigate the random effects, they mix a Poisson distribution with log-normal distribution; however, the approach may not account for the over-dispersion. In this paper, we estimate different versions of generalized standard distributions and zero-inflated models discussed by Harris et al. (2014) and Hilbe (2011).

We investigate how accounting for inflated zeros impacts the effects of moral hazard effects and adverse selection, which affects how individuals allocate their income to health care. We examine the importance of over-dispersion in the data, and select the model that results in more accurate predictions for the data from Riphahn et al. (2003). Comparing with Riphahn et al. (2003), the selected Zero-Inflated Negative Binomial2 (ZINB2) model estimates the impact of the adverse selection and moral hazard on health-care demand in a consistent way.

A number of papers have extended the application of zero-inflated models. Greene (1994) considers the zero-inflated negative binomial (ZINB). As Ainsworth (2007) argues, ZINB model is used by Neal and Gaher (2006) to study drug use issues among college students; Gupta et al. (1996) and Famoye and Singh (2006) apply a zero-inflated Generalize Poisson model to study frequentist setting. Ainsworth (2007) points out that Zero-inflated models have been further developed, in Ecology, by Ridout et al. (1998), Martin et al. (2005) and Kuhnert et al. (2005) to explain different kinds of zeros: those that are structural as well as those that depend on the study. Other papers, for example, Cohen (1960), Johnson et al. (2005) are focussing on underscore zeros, while Melkersson and Rooth (2000), Li et al. (2003) are focussing on situations where data have inflated zero.[1]

The outline of the paper is as follows. Section 2 discusses the different methodologies that are employed in this paper. Section 3 describes the data used in the

[1]As will be discussed in following sections, recently various extensions of zero-inflated models have been emphasised by Harris et al. (2014) and are incorporated in STATA. More details about some of these models are discussed in Hilbe (2011). While, STATA is not able to provide panel data estimates for zero-inflated models, LIMDEP is able to estimate fixed effect and random effect models in this context.

analysis, and Sect. 4 provides the results of different model specifications. Section 5 evaluates the predictions of the employed models, and Sect. 6 reviews robustness checks. Section 7 concludes.

2 Methodology

2.1 Discussion of Over Dispersion

In what follows we discuss the importance of properly accounting for over dispersion when it is present in count data models such as the ones used to model doctor and hospital visits. As Hilbe (2011) discusses, omitted variable, the existence of outliers, or clustering that results in correlation between responses can cause over-dispersion. Its presence in count data models raises the Pearson statistics adjusted with degree of freedom above one.[2] The uncontrolled over-dispersion may results in unreliable hypothesis test. In what follows we present how over-dispersion can be taken into account for count data models through mixing the Poisson distribution with other distributions.

2.2 Count Data Models

A Poisson model with equal mean and variance, $E(Y_i) = V(Y_i) = \mu_i$, has no power in dealing with over-dispersion. To make it more flexible, the model can be augmented with other distributions. This is done by relating its mean to an individual unobserved effect (u_i). We can obtain different extensions of the Poisson model depending on how we specify the distribution for u_i. Appendix B, Table 21, provides an extensive discussion about these extensions. A Generalized Poisson (GP) model can also accommodate both over-dispersion and under-dispersion. We use these distributions to discuss the robustness of the results in the analysed data.

2.3 Zero-Inflated Count Models

Data with more zeros than what we expect from a particular distribution may be a suspect of over-dispersion. Zero-inflated Poisson (ZIP) model and zero-inflated negative binomial (ZINB) models adjust for excessive zeros in the response. Hilbe

[2]See Hilbe 2011, chapter "The Contribution of Productivity and Price Change to Farm-level Profitability: A Dual Approach Analysis of Crop Production in Norway".

(2014) discuss different versions of zero-inflated models. The mixture framework of these distributions can explain more of over-dispersion in the data.

Appendix B, Table 22, presents different zero-inflated distributions that are used as robustness checks for our analysis. Following Hilbe (2011), a Vuong (1989) test[3] for non-nested models is used to compare the suitability of a zero-inflated distribution against its standard distribution as in Table 21. If Vuong test is positive and significant, the zero-inflated model is preferred to its corresponding standard one. With negative and significant value, the standard model is the selected one. With a non-significant Vuong test, none of them is preferred to the other one. As an additional check, we look at the predictability of different model specifications.

3 Data Description

We use the same data as in Riphahn et al. (2003) that is "the first twelve annual waves (1984 through 1995) of the German Socioeconomic Panel (GSOEP) which surveys a representative sample of East and West German households". The data set is downloadable from the web site of Journal of Applied Econometrics.[4] The data is restricted to individuals aged between 25 and 65. Table 11 presents the descriptive statistics of the dependent variable by gender.[5] Following Riphahn et al. (2003), the dependent variables are defined as "the number of visits to a doctor within the last quarter prior to the survey, and the number of inpatient hospital visits with at least one night spent in the hospital within a given calendar year".

Table 1 shows the presence of inflated-zeros in both hospital visits and doctor visits for both genders. Around 92% and 44% of males did not visit a hospital and a doctor. For females, the shares of zero hospital and doctor visit are around 90% and 30%, respectively. The abundance of zeros in both kinds of visits suggests that zero-inflated distributions might be better options rather than their standard versions for the purpose of examining doctor and hospital demands. Since the frequency of zeros for doctor visits is less than for hospital visits, this paper focuses on the demand equation for doctor visits. If the results for this equation confirm the superiority of zero-inflated distributions over their standard versions, the results can be extended to demand for hospital visits as well. Among the explanatory variables, Riphahn et al. (2003) consider two different dummy variables for two types of insurance: whether an individual has public insurance or not, and, if yes whether he or she has an add-on insurance policy, which is an optional policy to cover some other costs. They argue that 90% of German people have mandatory insurance policy with only 1% without any insurance.

[3] Vuong test is a likelihood ratio based test for selecting a specific model among non-nested models.

[4] See: http://qed.econ.queensu.ca/jae/

[5] For more detail about the data see Table 11 in the appendix as well as Riphahn et al. 2003.

Table 1 Dependent variables: Descriptive statistics

	(Share of total observation, %)			
	Hospital visit		Doctor visit	
Value	Males	Females	Males	Females
0	92.21	90.18	44.05	29.51
1	6.18	7.88	13.82	13.17
2	1.09	1.28	11.63	13.42
3	0.15	0.27	8.48	11.49
04-Sep	0.21	0.25	15.29	21.83
10 and more	0.16	0.14	6.73	10.58
Mean	0.128	0.15	2.63	3.79
Std dev.	0.93	0.83	5.21	6.11
Median	0	0	1	2
N	14,243	13,083	14,243	13,083

Source, German Socioeconomic Panel (1984–1995)

All the explanatory variables are the same as Riphahn et al. (2003); see Table 11 in the appendix.

4 Discussion of the Results

4.1 Panel Data Models

On the demand side, Riphahn et al. (2003) assume a bivariate model for the demands of doctor and hospital visits. These demands follow a Poisson distribution, and the unobservable heterogeneity and error terms follow lognormal and bivariate normal distributions, respectively:

$y_{itg} \sim Po(\mu_{itg}) \quad g = 1, 2$ (with 1 for doctor visits and 2 for hospital visits).

$\ln(\mu_{it1}) = \beta' x_{it1} + u_{i1} + \varepsilon_{it1}$; $u_{i1} \sim N\left(0, \sigma_{u1}^2\right)$; $(\varepsilon_{it1}, \varepsilon_{it2}) \sim N_2\left(0, 0, \sigma_{\varepsilon 1}^2, \sigma_{\varepsilon 2}^2, \rho\right)$; $E[u_{ig}u_{jh}] = 0$ *if* $i \neq j \vee g \neq h$.

$\ln(\mu_{it2}) = \beta' x_{it2} + u_{i2} + \varepsilon_{it2}$; $u_{i2} \sim N\left(0, \sigma_{u2}^2\right)$; $E[\varepsilon_{itg}u_{jh}] = 0 \quad \forall\, i, t, g, j, h$; $E[\varepsilon_{itg}\varepsilon_{jsh}] = 0$ *if* $t \neq s\ i \neq j \vee g \neq h$

To integrate out the unobserved heterogeneity u_{ig}, a Gauss- Hermite approximation was used, while to integrate the distribution of cross-equation errors (ε_{itg}) a modified Gauss-Legendre approach was applied.

Riphahn et al. (2003) use the public insurance dummy to check for moral hazard and the add-on insurance dummy to check for adverse selection. The results indicate no evidence of moral hazard for demands for doctor visits and hospital visits: the coefficient of public insurance dummy are statistically insignificant and negative for male's hospital demand. Their model estimates positive coefficients for add-ons for both demands, though it is statistically significant only for males' hospital demand

and confirming adverse selection for it. Also they found that self-employed females and males have fewer visits to doctors than other employees.

4.2 Random Effects Model

As mentioned above, Riphahn Overall (2003) do control for unobserved effects by mixing Poisson distributions, however, their model lacks accounting for the inflated zeros in the data. Consequently, by taking into account this issue one can provide more reliable estimated parameters for add-on and public insurance variables.

To reconcile our analysis with Riphahn et al. (2003), we first estimate random effects models for doctor visits using a Gaussian distribution and a Gamma distribution to account for the unobserved heterogeneity. For the purpose of comparing our model with their model, we focus on the estimated coefficients for public insurance and add-on insurance dummies.

Table 2 reports the doctor visits' results for both females and males. Based on AIC and BIC criteria, we see that Gamma distribution is a better choice for the data. Also, with Gamma distribution the coefficient of public insurance is positive and statistically significant for both females and males while in the case of Gaussian distribution both of them are positive but insignificant. For both models, the coefficient of add-on is negative but not significant. The results show that,

Table 2 Random effect model with Gaussian and Gamma distributions for the unobserved heterogeneity term

	Males		Females	
	RE Normal	RE Gamma	RE Normal	RE Gamma
Doctor visit equation				
Public Insurance	0.106	0.103***	0.0638	0.0690*
	(0.0844)	(0.0388)	(0.0733)	(0.0360)
Add-on Insurance	−0.0334	−0.0340	−0.0260	−0.0317
	(0.103)	(0.0535)	(0.0897)	(0.0456)
Lnsig2u	0.0138		−0.248***	
	(0.0393)		(0.0391)	
Lnalpha		−0.00860		−0.277***
		(0.0293)		(0.0286)
Observations	14,243	14,243	13,083	13,083
AIC	65802.1774	65713.7349	70856.0950	70728.1257
BIC	65976.1498	65887.7074	71028.1136	70900.1443
Log lik.	−32878.1	−32833.9	−35405.0	−35341.1

Source, German Socioeconomic Panel (1984–1995)
Standard errors in parentheses
*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

although it seems that Gaussian distribution is more flexible than Gamma, NB2 (a mixture of Poisson and Gamma) is a better option for this data.

4.3 Standard Count Models

Following Greene (2008), we use pooled data to select the best model between standard models. Table 3 presents the results for males' visits to doctors when employing the models in Table 21: Poisson, Negative Binomial 1 (NB1), Negative

Table 3 Standard distributions for doctor visit for males

	Poisson	NB1	NB2	Gen_Possion	NBFamoy	GNBWaring
Doctor visit equation visiequation						
Public insurance	0.100	0.0607	0.0934	0.0595	0.0934	0.0578
	(0.0702)	(0.0539)	(0.0635)	(0.0549)	(0.0635)	(0.0577)
Add-on insurance	0.0666	0.139*	0.0551	0.144*	0.0551	0.154*
	(0.102)	(0.0777)	(0.0948)	(0.0791)	(0.0948)	(0.0844)
Constant	2.771***	2.776***	3.149***	2.780***	3.710***	2.929***
	(0.336)	(0.254)	(0.329)	(0.258)	(0.330)	(0.273)
Lndelta		1.581***				
		(0.0365)				
Lnalpha			0.561***			
			(0.0270)			
Atanhdelta				0.726***		
				(0.0115)		
Lnphim1					−17.76***	
					(3.253)	
Lntheta					−0.561***	
					(0.0270)	
Lnrhom2						0.783***
						(0.0981)
Lnk						2.303***
						(0.130)
Observations	14,243	14,243	14,243	14,243	14,243	14,243
AIC	85593.4779	54865.9120	55006.8616	54700.9022	55008.8616	54528.6162
BIC	85759.8863	55039.8845	55180.8341	54874.8747	55190.3981	54710.1527
Log lik.	−42774.7	−27410.0	−27480.4	−27327.5	−27480.4	−27240.3
Dispersion	6.67597	constant	1.998817			

Source, German Socioeconomic Panel (1984–1995)
*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Binomial 2 (NB2), Generalized Poisson (GPoisson), Negative Binomial Famoye (NB Famoye), and Negative Binomial Waring (NB Waring).

The results for Poisson, NB1 and NB2 are the same as in Greene (2008). Based on the dispersion criteria which has a high dispersion value of 6.67, Poisson distribution is not suitable. NB2 defeats Poisson by reducing considerably the value of dispersion to 1.99. This can also be confirmed by $\ln(\alpha) = 0.561$ which measures the logarithm of the dispersion parameter (α) and based on the likelihood ratio test is statistically significant. The same conclusion is obtained for other distribution: NB1, Generalized Poisson, NB Famoye and NB. Using AIC and BIC criteria, the NB Waring model is the best model for these data. It is followed by the Generalized Poisson, NB1, NB2, and NB Famoye.

NB Waring model estimates a positive and significant coefficient for the add-on insurance. In addition this parameter is positive in all models but only significant in NB1 and Generalized Poisson and NB Waring models. The estimated parameter for public insurance is positive in all the models but statistically insignificant. Table 12 in the appendix A reports all parameter estimates.

Table 13 in the appendix A, shows the same results for females. The results for dispersion and the ranking of the best models are the same as for males. Still, NB Waring is the best one and NB2 is in the second rank. The only difference is observed in add-on's estimated parameter. This parameter is positive, and, as for males, statistically significant for NB1 and generalized Poisson but not for NB Waring. For public insurance, all the models provide positive values except for the Poisson model, which is a completely unreliable model, where it is statistically significant. This can be viewed as evidence of over-dispersion, leading to underestimation of standard errors, making the coefficient statistically significant.

To conclude this analysis, we can state that even if we ignore the zero-inflated nature of the data, we can designate NB family and Generalized Poisson as better choices than the simple Poisson model.

Finally, Table 4 provides results for heterogeneous NB2. For this model, all the explanatory variables are used to explain its dispersion parameter. In comparison, AIC and BIC criteria indicates that this model is better than the simple Poisson for both males and females. Also, based on AIC criteria, this model has the lowest value compared to other specifications. In this model only the coefficient of public insurance for females is statistically significant.

4.4 Zero-Inflated Models (Pooled Data)

Following Greene (2008) again, we use pooled data to select the best model among zero-inflated models introduced in Table 22. Since for the zero inflated models it is necessary to specify the inflation function, all the explanatory variables are covariates in this function.

Table 5 provides estimation results related to zero-inflated models for males. Based on the positive and statistically significant values of Vuong statistics (the

Table 4 Heterogeneous NB2 for doctor visit for males and females

	Hete_NB2_Male	Hete_NB2_Female
Doctor visit equation		
Public insurance	0.0940	0.105*
	(0.0652)	(0.0599)
Add-on insurance	0.0427	0.0347
	(0.0927)	(0.0754)
Constant	2.977***	2.874***
	(0.318)	(0.272)
Lnalpha		
Public insurance	0.0188	0.0190
	(0.0985)	(0.102)
Add-on insurance	−0.397***	−0.495***
	(0.153)	(0.142)
Constant	−0.839*	−1.217***
	(0.452)	(0.453)
Observations	14,243	13,083
AIC	54493.3321	60278.6270
BIC	54826.1490	60607.7061
Log lik.	−27202.7	−30095.3

Source, German Socioeconomic Panel (1984–1995)
*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

test for non-nested models), there is a strong evidence to prefer the zero-inflated models to their corresponding standard models. Also, add-on contributes to the over-dispersion in the data as it is significant in the zero-inflated function. The Zero-Inflated Negative Binomial (ZINB) Waring model has the lowest AIC and BIC values followed by ZINB2 ZINB-Famoye and zero-Inflated Poisson (ZIP). The statistically significant estimated parameters related to dispersion in ZINB Waring and ZINB2 indicate that zero-inflated Poisson is not a good choice for these data and its significant coefficient for add-on is not reliable.

The estimated parameters for add-on are negative for all of the models but statistically insignificant except for ZIP. Moreover, the public insurance coefficient is positive and statistically significant ZIP and ZINB2.

Table 6 provides the results of zero-inflated models for females. The results for females are similar to those for males. In the case of females, the coefficient for public insurance is statistically positive for ZINB2 model.

Table 5 Zero-Inflated models for males

	ZIP	ZINB2	ZINBFamoye	ZINBWaring
Doctor visit equation				
Public insurance	0.0794***	0.0971*	0.0899	0.0565
	(0.0247)	(0.0571)	(0.0560)	(0.0623)
Add-on insurance	−0.0839*	−0.0388	−0.0694	−0.00574
	(0.0430)	(0.0962)	(0.0933)	(0.101)
Constant	2.502***	2.567***	−5.078	2.598***
	(0.108)	(0.263)	(154.3)	(0.291)
Inflate equation				
Public insurance	−0.0226	0.0342	0.00727	−0.00330
	(0.0755)	(0.162)	(0.124)	(0.148)
Add-on insurance	−0.423***	−0.651	−0.590*	−0.637*
	(0.157)	(0.446)	(0.316)	(0.387)
Constant	−3.718***	−6.989***	−5.710***	−4.933***
	(0.371)	(0.935)	(0.670)	(0.772)
Lnalpha		0.154***		
		(0.0303)		
Lnphim1			6.560	
			(154.4)	
Lntheta			7.652	
			(154.3)	
Lnrhom2				0.866***
				(0.0550)
Constant				0.897***
				(0.103)
Observations	14,243	14,243	14,243	14,243
AIC	70905.8533	54536.9885	54383.8199	54168.3164
BIC	71238.6702	54877.3694	54731.7649	54516.2613
Log lik.	−35408.9	−27223.5	−27145.9	−27038.2
Vuong_statistic	31.546871***	11.554577***	14.015931***	21.082627***

Source, German Socioeconomic Panel (1984–1995)
Standard errors in parentheses
*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

5 Model Evaluation

5.1 Distribution Comparisons

In this section, we compare the predictions of Poisson, zero-inflated Poisson (ZIP), NB2, and ZINB2 for doctor‘s visit with the corresponding actual distribution.[6]

[6]Since the predicted values might not be integers, we convert them to the closest integer.

Table 6 Zero-inflated models for females

	ZIP	ZINB2	ZINBFamoy	ZINBWaring
Doctor visit equation				
Public insurance	0.112***	0.0788*	0.0674	0.0595
	(0.0217)	(0.0466)	(0.0459)	(0.0495)
Add-on insurance	−0.0652*	−0.0146	−0.0333	0.0649
	(0.0337)	(0.0768)	(0.0758)	(0.0753)
Constant	2.307***	2.586***	−5.259	2.663***
	(0.0922)	(0.222)	(134.6)	(0.228)
Inflate equation				
Public insurance	−0.0857	−0.207	−0.149	−0.122
	(0.0855)	(0.194)	(0.145)	(0.165)
Add-on insurance	−0.427**	−0.913	−0.733	−0.459
	(0.176)	(0.887)	(0.476)	(0.416)
Constant	−4.329***	−8.360***	−6.726***	−5.556***
	(0.417)	(1.218)	(0.829)	(0.883)
Lnalpha		−0.0723***		
		(0.0258)		
Lnphim1			6.579	
			(134.7)	
Lntheta			7.879	
			(134.6)	
Lnrhom2				1.091***
				(0.0464)
Lnk				1.142***
				(0.0868)
Observations	13,083	13,083	13,083	13,083
AIC	79784.3595	60296.0349	60130.3157	59863.0979
BIC	80113.4386	60632.5930	60474.3529	60207.1351
Log lik.	−39848.2	−30103.0	−30019.2	−29885.5
Vuong_statistic	31.52***	8.80***	12.18 ***	12.67***

Source, German Socioeconomic Panel (1984–1995)
Standard errors in parentheses
*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Figure 1a and b compare the predicted frequency of the number of doctor visits by Poisson and ZIP with their actual frequencies. Both figures show that Poisson and zero-inflated Poisson models underestimate the zeros and overestimate the ones. Figure 1c compares the distributions results from the Poisson and zero-inflated Poisson models. We see that zero-inflated Poisson increases the estimated frequency of zeros by almost 40%, which is a substantial improvement in terms of prediction. We observe also some improvements in the reduction of the estimated one and two visits. Regarding four and more visits, both models are nearly the same.

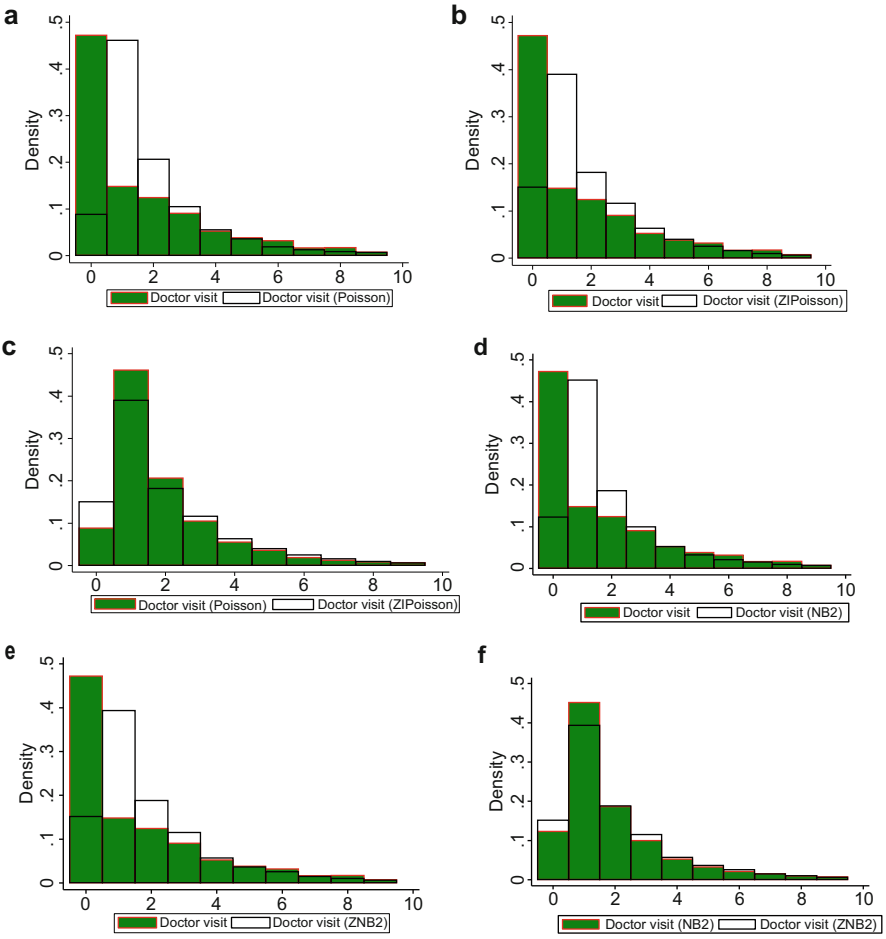


Fig. 1 Doctor visits (DV) and predicted DV by model specification (**a**) Doctor visits (DV) and predicted DV (by Poisson) (**b**) DV and predicted DV (by ZIPoisson) (**c**) Predicted DV by Poisson and ZIPoisson (**d**) DV and predicted DV (by NB2) (**e**) DV and predicted DV (by ZNB2) (**f**) Predicted DV by NB2 and ZNB2

Figure 1c, d and e compare the results for NB2 and zero-inflated NB2. The improvement in the number of zeros using zero-inflated NB2 is almost 20% compared with the standard NB2. This improvement is less than what we mentioned above comparing Poisson and zero-inflated Poisson. We can justify it since comparing with Poisson model the standard NB2 has more power in accounting for over-dispersion, as we discussed before. This can also be seen in Tables 14, 15, 16 and 17 of the appendix A. For example, standard Poisson distribution predicts correctly 838 zeros while standard NB2 predicts correctly 1170 zeros. With zero-inflated Poisson and zero-inflated NB2 these values increase to 1441 and 1446,

respectively. In addition, the zero-inflated models provide some improvements on the estimation of other numbers of doctor visits.

5.2 *Predicted Versus Realizations Comparisons*

Table 7 compares four models based on the maximum differences and mean absolute differences between predicted and actual counts. The results show that Poisson performs worst at predicting the 0 s, and NB2 and ZIP perform worst at predicting the 1 s, while ZINB2 is worst at predicting the 2 s. However, the maximum difference and mean of absolute differences are much lower for ZINB2 which means this model is the best one in terms of overall prediction. The Pearson statistic equals 193.429 for this model (the lowest of all models), which confirms it is the best model in terms of prediction (see Table 18 in the appendix A for more details).

Figure 2 presents the density comparison between actual and predicted probabilities. Again, we see that ZINB2 is superior to ZIP in predicting actual probabilities. Further, Fig. 3 plots the residuals from the tested models. Small residuals indicate a good fit, so the models with lines closest to zero should be considered as the suitable ones. We can see that the residuals line for ZINB2 is very close to zero when compared with the line of residuals for all the other models, confirming the results of all previous findings.

Finally, Table 8 provides tests for choosing the best model in terms of fit statistics such as AIC and BIC as well as Vuong statistic. The results also indicate that ZINB2 is the best model among the models under consideration.

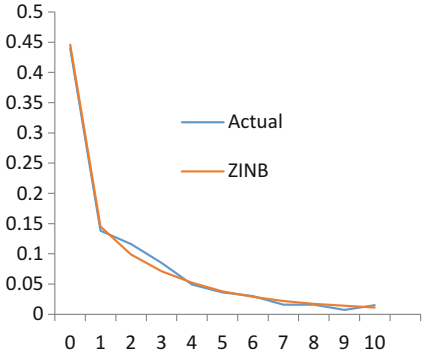
6 Robustness Checks

6.1 *Robustness Results on Pooled Sample*

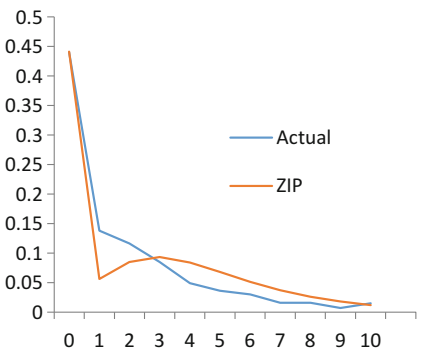
The results of previous sections show that zero-inflated NB2 could be considered a suitable model in terms of prediction for both female and male subsamples. The model predicts that public insurance has a positive and statistically significant

Table 7 Comparing the mean of observed and predicted count

Model	Maximum difference	At value	Mean |Diff|
Poisson	0.273	0	0.054
NB2	−0.042	1	0.009
ZIP	0.082	1	0.023
ZINB2	0.017	2	0.006



Actual probabilities versus Zero Inflated Negative Binomial



Actual probabilities versus Zero Inflated Poisson

Fig. 2 Density comparison between actual and predicted probabilities

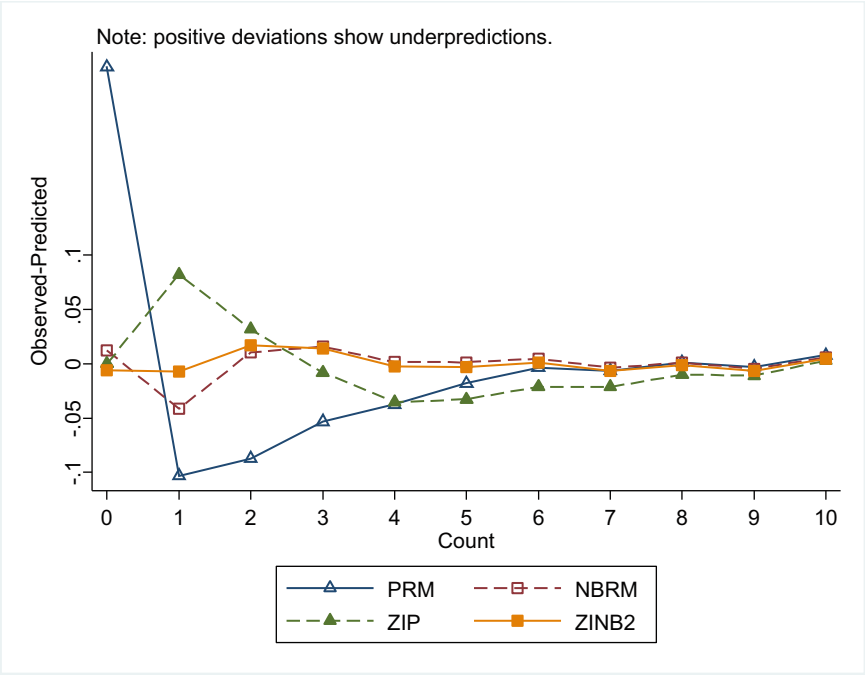


Fig. 3 Residual plots of PRM(Poisson Random Effect Model, NBRM (NB Random Effect Model), ZIP(Zero Inflated Poisson Model), ZINB(Zero Inflated NB Model)

Table 8 Tests of fit statistics

Poisson	**BIC = 85759.8**	**AIC = 85593.5**	**Prefer**	**Over**	**Evidence**
Vs NB2	BIC = 55180.8	Diff = 30579.1	NB2	Poisson	Very strong
	AIC = 55006.8	Diff = 30586.6	NB2	Poisson	
	LRX2 = 30588.6	Prob = 0.000	NB2	Poisson	P = 0.000
Vs ZIP	BIC = 71238.7	Diff = 14521.2	ZIP	Poisson	Very strong
	AIC = 70905.8	Diff = 14687.6	ZIP	Poisson	
	Vuong = 31.54	Prob = 0.000	ZIP	Poisson	P = 0.000
Vs ZINB2	BIC = 54877.4	Diff = 30882.5	ZINB	Poisson	Very strong
	AIC = 54536.9	Diff = 31056.5	ZINB	Poisson	
NB2	**BIC = 55180.8**	**AIC = 55006.8**	**Prefer**	**Over**	**Evidence**
Vs ZIP	BIC = 71238.7	Diff = − 16057.8	NB2	ZIP	Very strong
	AIC = 70905.8	Diff = − 15898.9	NB2	ZIP	
Vs ZINB2	BIC = 54877.4	Diff = 303.4	ZINB2	NB2	Very strong
	AIC = 54536.9	Diff = 469.8	ZINB2	NB2	
	Vuong = 11.55	Prob = 0.000	ZINB2	NB2	P = 0.000
ZIP	**BIC = 71238.7**	**AIC = 70905.8**	**Prefer**	**Over**	**Evidence**
Vs ZINB2	BIC = 54877.4	Diff = 16361.3	ZINB2	ZIP	Very strong
	AIC = 54536.9	Diff = 16368.8	ZINB2	ZIP	
	LRX2 = 16370.8	Prob = 0.000	ZINB2	ZIP	P = 0.000

coefficient for both subsamples, which has implications on the income effects of individuals.

Now, we check if we can get the same results as before by pooling males and females data. We add a dummy variable with value of 1 for females and value of 0 for males to the explanatory variables. Consequently, six different models for the whole sample are estimated: Poisson with Gaussian random effect, Poisson with Gamma random effect, NB2, NB Waring, zero-inflated NB2 as well as zero-inflated NB Waring.

Table 9 reports the estimation results for the six models. In all of them the estimated coefficients for females are positive and statistically different from zero which confirms our focusing on two separate samples for males and females in the previous sections. It also means that on average, *ceteris paribus*; females will demand more visits for doctors than males. Between the random effect models, the one with Gamma distribution for the unobserved heterogeneity performs best (both AIC and BIC are predicting the same result). However, all the pooled data models are preferred to random-effect models. Further, the Voung statistic used to compare non-nested models is positive and statistically significant for the zero-inflated models, meaning that zero-inflated models provide better predictions than their standard counterparts. Furthermore, based on AIC and BIC, the zero-inflated Waring model, though less accurate in convergence, is preferred to the zero-inflated NB2. Finally, all the models estimate a positive coefficient for public insurance and are statistically significant except for the Gaussian random effect model. The results confirms the existence of moral hazard.

Table 9 Full sample results

	Gaussian RE	Gamma RE	NB2	NBW	ZINB2	ZINB waring
Doctor visit equation						
Female	**0.377*****	**0.304*****	**0.354*****	**0.364*****	**0.183*****	**0.215*****
	(0.0291)	(0.0244)	(0.0279)	(0.0234)	(0.0194)	(0.0207)
Public insurance	0.0886	**0.0896*****	**0.1000****	**0.0746***	**0.0930****	**0.0654***
	(0.0557)	(0.0264)	(0.0458)	(0.0393)	(0.0361)	(0.0389)
Add-on insurance	−0.0299	−0.0327	0.0497	0.148***	−0.0193	0.0340
	(0.0665)	(0.0347)	(0.0613)	(0.0553)	(0.0597)	(0.0606)
Lnsig2u	−0.111***					
	(0.0277)					
Lnalpha		−0.136***	0.370***		0.0243	
		(0.0205)	(0.0189)		(0.0195)	
Lnrhom2				0.842***		0.984***
				(0.0652)		(0.0349)
Lnk				2.280***		1.043***
				(0.0923)		(0.0665)
Inflate equation						
Female					−1.216***	−0.830***
Public insurance					−0.0469	−0.0477
					(0.120)	(0.107)
Add-on insurance					−0.682**	−0.562**
					(0.341)	(0.269)
Observations	27,326	27,326	27,326	27,326	27,326	27,326
AIC	136878.2159	136666.2365	115861.5909	114914.2459	114881.7180	114051.5364
BIC	137075.3902	136863.4108	116058.7651	115119.6357	115267.8509	114445.8849
Log lik.	−68415.1	−68309.1	−57906.8	−57432.1	−57393.9	−56977.8
Vuong_statistic					15.94***	39.64***

Source, German Socioeconomic Panel (1984–1995)
Standard errors in parentheses
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

6.2 *Accounting for Correlation Between Doctor Visits and Hospital Visits Using the Bivariate Negative Binomial Model*[7]

We also jointly estimate both doctor visits and hospital visits using a bivariate NB model to account for potential correlation between the two events. As Riphahn et al. (2003) explain, doctor visits and hospital visits are positively correlated. However, this correlation should be identified and tested.

In the males sample we find a correlation between doctor visits and hospital visits of 0.1477. Using only non-zero values, this correlation reduces to 0.1138. Cameron and Trivedi (2013) show how to construct a statistic for testing the independency between two counts specific regressions (for doctor visits and hospital visits). The calculated test statistic for our sample are 9.47, 1.05, 0.05, 0.94, respectively for $(j,k) = (1,1),(1,2),(2,1),(2,2)$. Only the first value is statistically different from zero (with p-vale equal to 0.002). This shows that independency can only be rejected for the first statistics suggesting some evidence of weak dependency between the two count variables. This is in contrast to what Riphahn et al. (2003) expected. Motivated by the first test value, however, a bivariate NB2 model is estimated for males using pooled panel data (see Table 10). Here the parameter α can capture both overdispersion as well as correlation between unobserved heterogeneity[8].

For females, the correlation between doctor visits and hospital visits for all data realizations (including the zeros) is 0.125 and when we look at only the positive values we get a correlation of 0.079. The independency test statistic for females

Table 10 Bivariate negative binomial 2 for doctor visits and hospital visits

	Males		Females	
	Doctor visit	Hospital visit	Doctor visit	Hospital visit
Public Insurance	0.0958*	−0.173	0.105*	−0.144
	(0.0551)	(0.235)	(0.0545)	(0.312)
Add-on Insurance	0.0605	0.550	0.0311	0.0501
	(0.0863)	(0.351)	(0.0648)	(0.164)
Alpha	1.698***		1.169***	
	(0.0367)		(0.0242)	
Observations	14,243		13,083	
Log Like	−33090.1		−36174.5	

Source, German Socioeconomic Panel (1984–1995)
Standard errors in parentheses

[7]Codes for Bivariate NB2 model are found in Cameron, C., and Trivedi, P. (2013) (see page 336).

[8]For the estimating Bivariate NB2 by ML, initial values we find by estimating non-linear seemingly unrelated regression (NLSUR) and assuming initial value for α equal to 2. For first stage, correlation between doctor visits and hospital visits for males and females are estimated as 0.125 and 0.078, respectively.

are 38.56, 0.087, 4.81, 0.225 corresponding to $(j,k) = (1,1),(1,2),(2,1),(2,2)$, respectively. Based on the results, the first and second values of the test statistic are significant at 0.05 level. Motivated by the first two test values, a bivariate NB2 is estimated for female. The results are presented in Table 10.

We see that the α parameters are significant in both male and female bivariate models. This result might confirm the previous results that NB2 based models are a better distribution for explaining real data, rather than Poisson distribution. Here we find that the estimated public insurance parameters for doctor visits are positive and statistically significant for both males and females while for hospital visits are negative and insignificant. Although estimated add-on parameters are positive for all males and females in two equations, they are statistically insignificant.

7 Conclusion

The high share of zeros for a dependent variable in a count data regression model can severely increase the over-dispersion issue and lead to unreliable estimators. We show that the German Socioeconomic Panel (1984–1995) used by Riphahn et al. (2003) for the demand of doctor visits suffers severely from over-dispersion issue and their estimation based on standard distributions might not be reliable. Results based on standard distributions are close to Riphahn et al. (2003) and, overall, there is not enough evidence for moral hazard and adverse selection except for Waring NB2, which presents a positive effect from adverse selection on the number of doctor visits. However, this result might also not be reliable due to the over-dispersion that resulted in the high share of zeros in the data. Overall we show that within the class of random effect models, the model with a Gamma distribution for unobserved heterogeneity is more suitable than the one assuming Gaussian distribution for unobserved heterogeneity.

Vuong test (1989) rejects the standard distributions in the favour of their corresponding zero-inflated distributions. This means that over-dispersion due to the high share of zeros in the data cannot be explained by any complex and/or flexible mixture of Poisson distributions such as Negative Binomial 2, Generalized Poisson, Negative Binomial Famoy, Generalized Negative Binomial Waring models. All of these are inferior to the zero-inflated distributions models. Between zero-inflated distributions, ZINB Waring model has the lowest AIC and BIC values followed by ZINB2 Famoye and ZINB2. However, when ranking the predicted probabilities, ZINB2 model produces the closest probabilities to the actual probabilities. A pooled (male-female) sample estimation provides the same results as those obtained from subsample estimations.

In contrast to Riphahn et al. (2003), most of the zero-inflated distribution models predict a negative but insignificant coefficient for add-on insurance for both male and female subsamples. This results indicate a weak sign of adverse selection. We find a positive coefficient for public insurance in all the estimated models, but only

statistically significant for ZINB2, the best model, for both genders. This result confirms the existence of moral hazard in the insurance market.

Although we show that the correlation between the demand for doctor and hospital is weak in the data, our bivariate NB model finds a positive and significant coefficient for public insurance and a positive and significant for add-on insurance for doctor's visit. This is also in contrast with the results by Riphahn et al. (2003), who do not find any significant coefficient for the public insurance for doctor's visit in their bivariate model. Overall, our results find a strong evidence for moral hazard for visiting more doctors. The results provide significant evidence of how considering the over-dispersion nature of the data in the estimation process can provide more precise estimations and reveal a better understanding about health demand components.

Additional Tables

Table 11 Summary Statistics

Variables	Description	Males[a]			Females[a]		
Docvis	Number of doctor visits in last three months	2.626		(5.21)	3.791		(6.11)
Hos	Number of hospital visit last year	0.128		(0.93)	0.150		(0.83)
Age	Age	42.653		(11.27)	44.467		(11.32)
Hsat	Health satisfaction code 0 (low)-10 (high)	6.924		(2.25)	6.634		(2.33)
Handdum	Person is handicapped (0/1)	0.227		(0.42)	0.200		(0.40)
Handper	Percentage degree of handicap	8.134		(20.33)	5.791		(17.96)
Married	Person is married (0/1)	0.765		(0.42)	0.752		(0.43)
Educ	Years of schooling	11.729		(2.44)	10.876		(2.11)
Hhninc	Monthly household net income ($\times 10^{-3}$)	3.591		(1.74)	3.445		(1.80)
Hhkids	Children below age 16 in household (0/1)	0.413		(0.49)	(0.392)		(0.49)
Self	Person is self-employed (0/1)	0.086		(0.28)	0.037		(0.19)

(continued)

Table 11 (continued)

Variables	Description	Males[a]			Females[a]		
Self	Person is self-employed (0/1)	0.086		(0.28)	0.037		(0.19)
Beamt	Person is civil servant (0/1)	0.118		(0.32)	0.028		(0.16)
Bluec	Person is blue collar worker (0/1)	0.340		(0.47)	0.139		(0.35)
Working	Person is employed (0/1)	0.850		(0.36)	0.488		(0.50)
Public Insurance	Person is insured in public health insurance (0/1)	0.861		(0.35)	0.913		(0.28)
Add-on Insurance	Person is insured in add-on insurance (0/1)	0.018		(0.13)	0.020		(0.14)
d85	Year = 1985 (0/1)						
d86	Year = 1986 (0/1)						
d87	Year = 1987 (0/1)						
d88	Year = 1988 (0/1)						
d91	Year = 1991 (0/1)						
d94	Year = 1994 (0/1)						
N	Number of observations		14,243			13,083	

Source: German Socioeconomic Panel (1984–1995)
[a]mean, standard deviation in parentheses

Table 12 Standard distributions for doctor visit for males (complete table)

	Poisson	NB1	NB2	Gen_Possion	NBFamoy	GNBWaring
Doctor visit						
Age	−0.0239	−0.0477***	−0.0398***	−0.0496***	−0.0398***	−0.0533***
	(0.0164)	(0.0114)	(0.0153)	(0.0114)	(0.0153)	(0.0120)
Age2	0.369**	0.634***	0.547***	0.659***	0.547***	0.706***
	(0.184)	(0.129)	(0.176)	(0.130)	(0.176)	(0.137)
Hsat	−0.225***	−0.189***	−0.239***	−0.188***	−0.239***	−0.203***
	(0.00767)	(0.00585)	(0.00745)	(0.00587)	(0.00745)	(0.00657)
Handdum	0.0690	0.0229	−0.0209	0.0183	−0.0209	0.0111
	(0.0537)	(0.0378)	(0.0503)	(0.0373)	(0.0503)	(0.0397)
Handper	0.00286**	0.00414***	0.00661***	0.00430***	0.00661***	0.00505***
	(0.00121)	(0.000848)	(0.00116)	(0.000835)	(0.00116)	(0.000917)
Married	0.0583	0.130***	0.0658	0.135***	0.0658	0.139***
	(0.0606)	(0.0408)	(0.0535)	(0.0409)	(0.0535)	(0.0432)
Educ	−0.0235***	−0.00955	−0.0262***	−0.00833	−0.0262***	−0.00971
	(0.00873)	(0.00672)	(0.00910)	(0.00688)	(0.00910)	(0.00725)

(continued)

Table 12 (continued)

	Poisson	NB1	NB2	Gen_Possion	NBFamoy	GNBWaring
Bhninc	−0.0000222*	−0.00000788	−0.0000192*	−0.00000746	−0.0000192*	−0.00000835
	(0.0000121)	(0.00000853)	(0.0000105)	(0.00000868)	(0.0000105)	(0.00000911)
Bhkids	−0.0760	−0.0743**	−0.0844*	−0.0766**	−0.0844*	−0.0792**
	(0.0518)	(0.0341)	(0.0470)	(0.0343)	(0.0470)	(0.0362)
Self	−0.211**	−0.244***	−0.218***	−0.253***	−0.218***	−0.265***
	(0.0847)	(0.0616)	(0.0784)	(0.0628)	(0.0784)	(0.0656)
Beamt	0.0914	0.0278	0.0841	0.0273	0.0841	0.0254
	(0.0809)	(0.0623)	(0.0766)	(0.0630)	(0.0766)	(0.0664)
Bluec	0.0178	−0.00948	0.0371	−0.0116	0.0371	−0.00956
	(0.0486)	(0.0374)	(0.0458)	(0.0379)	(0.0458)	(0.0398)
Working	−0.0554	0.0126	−0.0155	0.0175	−0.0155	0.0172
	(0.0668)	(0.0465)	(0.0596)	(0.0465)	(0.0596)	(0.0490)
Public Insurance	0.100	0.0607	0.0934	0.0595	0.0934	0.0578
	(0.0702)	(0.0539)	(0.0635)	(0.0549)	(0.0635)	(0.0577)
Add-on Insurance	0.0666	0.139*	0.0551	0.144*	0.0551	0.154*
	(0.102)	(0.0777)	(0.0948)	(0.0791)	(0.0948)	(0.0844)
d85	0.0769	0.0615*	0.106*	0.0611*	0.106*	0.0669*
	(0.0563)	(0.0359)	(0.0546)	(0.0358)	(0.0546)	(0.0378)
d86	0.215***	0.156***	0.226***	0.155***	0.226***	0.163***
	(0.0597)	(0.0365)	(0.0581)	(0.0365)	(0.0581)	(0.0386)
d87	0.113	0.0967**	0.123**	0.0983**	0.123**	0.104**
	(0.0690)	(0.0439)	(0.0613)	(0.0433)	(0.0613)	(0.0458)
d88	0.0530	0.111***	0.0670	0.110***	0.0670	0.115***
	(0.0558)	(0.0360)	(0.0544)	(0.0361)	(0.0544)	(0.0379)
d91	−0.00397	0.145***	−0.00366	0.152***	−0.00366	0.151***
	(0.0609)	(0.0373)	(0.0531)	(0.0374)	(0.0531)	(0.0393)
d94	0.247***	0.268***	0.244***	0.278***	0.244***	0.289***
	(0.0613)	(0.0407)	(0.0548)	(0.0409)	(0.0548)	(0.0430)
Constant	2.771***	2.776***	3.149***	2.780***	3.710***	2.929***
	(0.336)	(0.254)	(0.329)	(0.258)	(0.330)	(0.273)
Lndelta,		1.581***				
		(0.0365)				
Lnalpha			0.561***			
			(0.0270)			
Atanhdelta				0.726***		
				(0.0115)		
Lnphim1					−17.76***	
					(3.253)	
Lntheta					−0.561***	
					(0.0270)	

(continued)

Table 12 (continued)

	Poisson	NB1	NB2	Gen_Possion	NBFamoy	GNBWaring
Lnrhom2						0.783***
						(0.0981)
Lnk						2.303***
						(0.130)
Observations	14,243	14,243	14,243	14,243	14,243	14,243
AIC	85593.4779	54865.9120	55006.8616	54700.9022	55008.8616	54528.6162
BIC	85759.8863	55039.8845	55180.8341	54874.8747	55190.3981	54710.1527
Dispersion	6.67597		constant	1.998817		
Log lik.	−42774.7	−27410.0	−27480.4	−27327.5	−27480.4	−27240.3

Source, German Socioeconomic Panel (1984–1995)
Standard errors in parentheses
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 13 Standard distributions for doctor visit for females (complete)

	Poisson	NB1	NB2	Gen_Possion	NBFamoy	GNBWaring
Doctor visit						
Age	−0.0132	−0.0322***	−0.0312***	−0.0347***	−0.0321***	−0.0400***
	(0.0121)	(0.00943)	(0.0115)	(0.00947)	(0.0116)	(0.0112)
Age2	0.179	0.396***	0.373***	0.425***	0.382***	0.479***
	(0.138)	(0.107)	(0.131)	(0.107)	(0.132)	(0.127)
Hsat	−0.203***	−0.171***	−0.208***	−0.170***	−0.208***	−0.218***
	(0.00641)	(0.00507)	(0.00631)	(0.00506)	(0.00636)	(0.00609)
Handdum	0.138**	0.106**	0.113**	0.102**	0.111**	0.119**
	(0.0565)	(0.0450)	(0.0485)	(0.0434)	(0.0487)	(0.0480)
Handper	0.00241**	0.00254***	0.00436***	0.00254***	0.00457***	0.00418***
	(0.00108)	(0.000867)	(0.00106)	(0.000846)	(0.00110)	(0.000998)
Married	0.0272	0.0440	0.0282	0.0455	0.0284	0.0366
	(0.0408)	(0.0322)	(0.0385)	(0.0323)	(0.0386)	(0.0377)
Educ	0.0147	0.0138*	0.00773	0.0136*	0.00740	0.0121
	(0.00933)	(0.00724)	(0.00894)	(0.00728)	(0.00898)	(0.00873)
Bhninc	−0.0000206**	−0.0000111	−0.0000162*	−0.0000103	−0.0000161*	−0.0000128
	(0.00000948)	(0.00000740)	(0.00000951)	(0.00000746)	(0.00000955)	(0.00000916)
Bhkids	−0.134***	−0.108***	−0.124***	−0.108***	−0.124***	−0.122***
	(0.0416)	(0.0311)	(0.0376)	(0.0311)	(0.0375)	(0.0367)
Self	−0.218**	−0.223***	−0.242***	−0.229***	−0.244***	−0.280***
	(0.0978)	(0.0705)	(0.0875)	(0.0707)	(0.0872)	(0.0849)
Beamt	−0.0711	−0.00922	−0.0198	−0.00859	−0.0183	−0.0499
	(0.117)	(0.0848)	(0.128)	(0.0859)	(0.129)	(0.107)
Bluec	−0.0354	−0.0718*	−0.0401	−0.0772**	−0.0406	−0.0730
	(0.0555)	(0.0392)	(0.0497)	(0.0392)	(0.0495)	(0.0471)

(continued)

Table 13 (continued)

	Poisson	NB1	NB2	Gen_Possion	NBFamoy	GNBWaring
Working	0.0149	0.0247	0.0305	0.0264	0.0313	0.0363
	(0.0392)	(0.0294)	(0.0354)	(0.0295)	(0.0354)	(0.0347)
Public Insurance	0.131**	0.0790	0.0953	0.0715	0.0935	0.0787
	(0.0599)	(0.0489)	(0.0639)	(0.0499)	(0.0643)	(0.0598)
Add-on Insurance	0.0207	0.126*	0.0309	0.138**	0.0312	0.111
	(0.0888)	(0.0682)	(0.0769)	(0.0687)	(0.0769)	(0.0794)
d85	−0.0362	−0.0326	−0.0127	−0.0303	−0.0119	−0.0218
	(0.0473)	(0.0319)	(0.0449)	(0.0318)	(0.0450)	(0.0386)
d86	0.0941**	0.0837**	0.102**	0.0836**	0.102**	0.114***
	(0.0449)	(0.0329)	(0.0430)	(0.0328)	(0.0433)	(0.0383)
d87	−0.0843	−0.0750	−0.0531	−0.0690	−0.0515	−0.0701
	(0.0642)	(0.0485)	(0.0566)	(0.0471)	(0.0569)	(0.0529)
d88	−0.180***	−0.0677**	−0.176***	−0.0670**	−0.176***	−0.145***
	(0.0448)	(0.0315)	(0.0439)	(0.0315)	(0.0441)	(0.0384)
d91	−0.154***	0.0108	−0.138***	0.0202	−0.138***	−0.0688*
	(0.0456)	(0.0326)	(0.0441)	(0.0327)	(0.0442)	(0.0402)
d94	0.197***	0.186***	0.221***	0.191***	0.222***	0.252***
	(0.0481)	(0.0370)	(0.0464)	(0.0371)	(0.0466)	(0.0433)
Constant	2.547***	2.731***	3.024***	2.777***	3.184***	3.190***
	(0.282)	(0.224)	(0.273)	(0.227)	(0.276)	(0.267)
Lndelta		1.549***				
		(0.0349)				
Lnalpha			0.188***			
			(0.0259)			
Atanhdelta				0.711***		
				(0.0108)		
Lnphim1					−4.580***	
					(0.762)	
Lntheta					−0.133***	
					(0.0443)	
Lnrhom2						1.014***
						(0.113)
Lnk						0.283***
						(0.0764)
Observations	13,083	13,083	13,083	13,083	13,083	13,083
AIC	91844.4596	60731.5683	60570.6248	60521.2975	60569.0256	60307.5709
BIC	92008.9991	60903.5869	60742.6434	60693.3160	60748.5232	60487.0686
Log lik.	−45900.2	−30342.8	−30262.3	−30237.6	−30260.5	−30129.8
Dispersion	6.689348		1.487322			

Source, German Socioeconomic Panel (1984–1995)
Standard errors in parentheses
*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Table 14 Predicted and original numbers of doctor visits (Standard Poisson)

Doctor visit	0	1	2	3	4	5	6
0	838	3564	1089	437	165	77	35
1	183	976	424	205	85	36	19
2	104	721	421	789	87	63	22
3	56	441	289	158	103	74	25
4	19	225	162	111	60	38	23
5	12	135	124	72	56	42	18
6	7	109	92	64	57	26	23
7	3	60	49	36	22	15	5
8	3	52	64	30	26	14	13
9	1	22	15	16	16	7	5
Total	1226	6305	2729	1318	677	392	188

Table 15 Predicted and original numbers of doctor visits (ZI Poisson)

Doctor visit	0	1	2	3	4	5	6
0	1441	2931	981	502	205	96	44
1	294	845	378	239	100	48	24
2	167	643	380	211	103	62	39
3	94	390	256	175	114	79	39
4	34	206	143	112	65	46	29
5	21	122	112	68	66	43	29
6	17	92	75	79	54	29	29
7	10	50	38	46	23	14	9
8	8	44	54	39	25	18	12
9	3	18	18	18	16	9	8
Total	2089	5341	2427	1489	771	444	262

Table 16 Predicted and original numbers of doctor visits (Standard NB2)

Doctor visit	0	1	2	3	4	5	6
0	1170	3360	971	406	172	63	42
1	242	954	386	199	73	39	23
2	173	735	369	187	72	58	33
3	77	442	266	147	91	96	35
4	27	235	145	97	56	40	21
5	18	139	116	60	60	38	23
6	10	113	80	66	45	24	23
7	7	59	44	38	21	12	4
8	5	55	54	33	23	15	9
9	1	22	13	16	15	7	4
Total	1694	6114	2444	1249	628	365	217

Table 17 Predicted and original numbers of doctor visits (ZINB2)

Doctor visit	0	1	2	3	4	5	6
0	1446	2932	1020	478	184	94	41
1	289	860	385	238	85	40	27
2	171	647	387	216	85	55	41
3	97	388	266	177	102	68	44
4	35	210	145	109	64	39	27
5	20	126	110	81	53	40	28
6	18	95	82	72	45	36	23
7	8	54	39	42	23	13	7
8	8	47	52	41	24	16	10
9	3	18	9	19	12	14	5
Total	2095	5377	2495	1473	677	415	253

Table 18 (NB2) Predicted and actual probabilities

Count	Actual	Predicted	|Diff|	Pearson
0	0.440	0.428	0.012	4.992
1	0.138	0.180	0.042	136.630
2	0.116	0.106	0.011	15.150
3	0.085	0.069	0.016	52.273
4	0.049	0.048	0.001	0.617
5	0.036	0.034	0.001	0.800
6	0.030	0.025	0.005	13.029
7	0.016	0.019	0.004	9.693
8	0.016	0.015	0.001	0.983
9	0.007	0.012	0.005	25.303
10	0.015	0.009	0.006	55.167
Sum	0.948	0.944	0.103	314.635

Table 19 (ZINB2) Predicted and actual probabilities

Count	Actual	Predicted	|Diff|	Pearson
0	0.440	0.446	0.006	1.139
1	0.138	0.145	0.007	5.034
2	0.116	0.099	0.017	40.468
3	0.085	0.071	0.014	40.148
4	0.049	0.052	0.003	1.765
5	0.036	0.038	0.003	2.878
6	0.030	0.029	0.001	0.591
7	0.016	0.022	0.007	28.703
8	0.016	0.017	0.001	1.759
9	0.007	0.014	0.006	44.089
10	0.015	0.011	0.005	26.856
Sum	0.948	0.945	0.070	193.429

Table 20 (Poisson) Predicted and actual probabilities

Count	Actual	Predicted	|Diff|	Pearson
0	0.440	0.441	0.000	0.000
1	0.138	0.056	0.082	1724.211
2	0.116	0.085	0.032	169.678
3	0.085	0.093	0.008	10.143
4	0.049	0.084	0.035	211.840
5	0.036	0.068	0.033	222.485
6	0.030	0.051	0.021	124.089
7	0.016	0.037	0.021	175.386
8	0.016	0.026	0.010	55.840
10	0.015	0.012	0.003	9.078
Sum	0.948	0.971	0.256	2796.77

Model Specifications

To account for over-dispersion and deviations from $E(Y_i) = V(Y_i) = \mu_i$ in the Poisson distribution, a new distribution is obtained by adding an individual unobserved effect (u_i) to the log of the mean of the Poisson model, $\ln(mean_i) = \ln(\mu_i) + \ln(u_i)$. Thus, by defining different distributions for u_i, new versions of the Poisson distribution are created. Table 21 presents a list of those distributions, known as standard distributions in this paper, with their variances. A Gamma distribution for u_i, for example, gives a Negative Binomial 2 (NB2) distribution with mean μ_i and conditional variance $\mu_i + \alpha\mu_i{}^2$, with the constant parameter α controlling for heterogeneity or dispersion among individuals. The additional parameter p in the Power Negative Binomial (NB-P) distribution, introduced by Greene (2008), provides NB1 or NB2 distributions when $p = 1$ or $p = 2$, respectively. Also, the Heterogeneous NB2 model allows the heterogeneity explained by α in the NB2 distribution to be a function of the individual's characteristics (z_i), $\alpha = \exp(z_i\gamma)$. Thus, α can vary among individuals. In special case, where $\phi \to 1$, the variance of Famoye's (1995) distribution approaches to that of the NB. The Waring Negative Binomial distribution introduce by Irwin (1968) converges to NB if $k \to \frac{1}{\alpha}, \rho \to \infty$. Also, if $\delta = 0$, the GP distribution reduces to the usual Poisson distribution with parameter θ_i. (See Hilbe (2011, 2014) for more details)

Zero-inflated Count Models

As Hilbe (2011) discuss, the framework of zero-inflated models are based on separating zero outcomes and positive ones. The probability of zero outcomes results from the group of individuals who are not the subject of an event ($Q(0)$ for those who do not have physician to visit), and those who are the subject of the

Table 21 The list of standard distributions

Distribution	Variance
Poisson	μ (where $\ln\mu = X\beta$)
Negative Binomial 1 (NB1)	$\mu + \alpha\mu$ (where α is the dispersion parameter)
Negative Binomial 2 (NB2)	$\mu + \alpha\mu^2$
Generalized Negative Binomial (NB-P)	$\mu + \alpha\mu^p$
Heterogeneous Negative Binomial (NB-H)	$\mu + \alpha_i\mu^2$ (where $\alpha_i = z_i\gamma$)
Generalized Negative Binomial (Famoy)	$\theta\mu(1-\phi\mu)(1-\phi\mu)^{-3}$
Waring Negative Binomial (NBW)	$\mu + \mu\left(\frac{k+1}{\rho-2}\right) + \mu^2\left\{\frac{k+\rho-1}{k(\rho-2)}\right\}$
Generalized Poisson (GP)	$\frac{1}{(1-\delta)^2}\mu$

Table 22 Zero-inflated distributions

Zero-inflated Poisson (ZIP)
Zero-inflated Negative Binomial 1 (ZINB1)
Zero-inflated Negative Binomial 2 (ZINB2)
Zero-inflated Generalized NB (ZINB-P)
Zero-inflated Poisson Inverse Gaussian, (ZIPIG)
Zero-inflated Generalized Poisson, (ZIGP)
Zero-Inflated 3-parameter Waring NB (ZINBW)
Zero-inflated 3-parameter Famoye NB (ZINBF)

event but with zero outcome (P(0) for those who do not visit their physicians). The two part of the model is written as:The probability of a zero outcome for the system is given by[9]:

$$\Pr(y = 0) = Q(0) + \{1 - B(0)\}\Pr(0)$$

And the probability of a nonzero count is[10]:

$$\Pr(y = k; k > 0) = \{1 - B(0)\}\Pr(k)$$

A Probit or logit model estimates $Q(0)$ while one of the standared models in Table (1) estimates $\Pr(k), k = 0, 1, \ldots n$. The mixture model have more power in explaining over-dispersion in the data (see also Hilbe and Greene (2008)).

Table 22 presents different zero-inflated distributions that are used in the next sections for the purpose of estimation and comparison.

[9]Stata gives this probability using the command: predict f0, pr(0).

[10]Stata gives this probability using the command: predict fk, pr(k).

References

Ainsworth. (2007). *Zero-inflated spatial models: Web supplement*. Lecture note. http://people.math.sfu.ca/~lmainswo/

Amponsah, S. (2013). Adverse selection, moral hazard, and income effect in health insurance: The case of Ghana. *Bulletin of Political Economy, Tokyo International University, 14*, 35.

Asante, F., & Aikins, M. (2008). *Does the NHIS cover the poor*. Accra: Danida Health Sector Support Office.

Bago d'Uva, T., & Jones, A. (2009). Health care utilization in Europe: New evidence from the ECHP. *Journal of Health Economics, 28*, 265–279.

Bajari, P., Hong, H., & Khwaja, A. (2011). *A semiparametric analysis of adverse selection and moral hazard in health insurance contracts*. Working Paper.

Bajari, P., Dalton, C., Hong, H., & Khwaja, A. (2014). Moral hazard, adverse selection, and health expenditures: A semiparametric analysis. *The Rand Journal of Economics, 45*(4), 747–763.

Bundorf, M. K., Herring, B., & Pauly, M. (2005). *Health risk, income, and employment-based health insurance*. NBER Working Paper No. 11677.

Cameron, C., Trivedi, T., Milne, F., & Piggott, J. (1988). A microeconometric model of the demand for health care and health insurance in Australia. *Review of Economic Studies, Oxford University Press, 55*(1), 85–106.

Cameron, C., & Trivedi, T. (2013). *Regression analysis of count data*. NewYork: Cambridge University Press.

Cardon, J. H., & Handel, I. (2001). Asymmetric information in health insurance: Evidence from the national medical expenditure survey. *The Rand Journal of Economics, 32*(3), 408–427.

Chiappori, P., & Salanie, B. (2000). Testing for asymmetric information in insurance markets. *Journal of Political Economy, 108*(1), 56–78.

Cohen, A. C. (1960). Estimation in a truncated Poisson distribution when zeros and ones are missing. *Journal of the American Statistical Association, 55*, 342–348.

Famoye, F. (1995). Generalized binomial regression model. *Biometrical Journal, 37*, 581–594.

Famoye, F., & Singh, K. P. (2006). Zero-inflated generalized poisson regression model with an application to domestic violence data. *Journal of Data Science, 4*, 117–130.

Geil, P., Million, A., Rotte, R., & Zimmermann, K. F. (1997). Economic incentives and hospitalization in Germany. *Journal of Applied Econometrics, 12*(3), 295–311.

Greene, W. H. (1994). *Some accounting for excess zeros and sample selection in poisson and negative binomial regression models*. (Working Paper EC-94-10): Department of Economics, New York University.

Greene, W. (2008). Functional forms for the negative binomial model for count data. *Economic Letters, 99*(3), 585–590.

Gupta, P. L., Gupta, R. C., & Tripathi, R. C. (1996). Analysis of zero-adjusted count data. *Computational Statistics and Data Analysis, 23*, 207–218.

Harris, T., Hilb, J., & Hardin, J. (2014). Modeling count data with generalized distributions. *The Stata Journal, 14*(3), 562–579.

Hilbe, J. (2011). *Negative binomial regression* (2nd ed.). NewYork: Cambridge University Press.

Hilbe, J. (2014). *Modeling count data*. Cambridge University Press.

Irwin, J. O. (1968). The Generalized Waring Distribution Applied to Accident Theory. *Journal of the Royal Statistical Society. Series A, 131*(2), 205–225

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions* (3rd ed.). New York: Wiley.

Keane, M., & Stavrunova, O. (2016). Adverse selection, moral hazard and the demand for Medigap insurance. *Journal of Econometrics, 190*(1), 62–78.

Kirigia, J. M., Sambo, L. G., Nganda, B., Mwabu, G. M., Chatora, R., & Mwase, T. (2005). Determinants of health insurance ownership among South African women. *BMC Health Services Research, 5*(1), 1.

Kuhnert, P. M., Martin, T. G., Mengersen, K., & Possingham, H. P. (2005). Assessing the impacts of grazing levels on bird density in woodland habitat: A Bayesian approach using expert opinion. *Environmetrics, 16*, 717–747.

Li, T., Trivedi, P. K., & Guo, J. (2003). Modeling response bias in count: A structural approach with an application to the national crime victimization survey data. *Sociological Methods and Research, 31*, 415–545.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., & Possingham, H. P. (2005). Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. *Ecology Letters, 8*, 1235–1246.

Marvasti, A. (2014). An estimation of the demand and supply for physician services using a panel data. *Economic Modelling, 43*, 279–286.

Melkersson, M., & Rooth, D. (2000). Modeling female fertility using inflated count data models. *Journal of Population Economics, 13*, 189–203.

Neal, S., & Gaher, R. (2006). Risk for marijuana-related problems among college students: An application of zero-inflated negative binomial regression. *American Journal of Drug and Alcohol Abuse, 32*, 41–53.

Pauly, M. V. (1968). The Economics of Moral Hazard. *American Economic Review, 58*(3), 531–537.

Powell, D. (2014). *Estimation of quantile treatment effects in the presence of covariates*. Unpublished manuscript.

Powell, D., & Goldman, D.. (2016). Disentangling Moral Hazard and Adverse Selection in Private Health Insurance, NBER Working Paper No. 21858.

Ridout, M., Demetrio, C., & Hinde, J. (1998). 'Models for count data with many zeros'. *International biometric conference*. Cape Town, 1998.

Riphahn, T., Wambach, A., & Million, A. (2003). Incentive effects in the demand for health care: A behavioral panel count data estimation. *Journal of Applied Economics, 18*, 387–405.

Rothschild, M., & Stiglitz, J. E. (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics, 90*(4), 629–649.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57*, 307–333.

Wolfe, J. R., & Goddeeris, J. H. (1991). Adverse selection, moral hazard, and wealth effects in the Medigap insurance market. *Journal of Health Economics, 10*(4), 433–459.

Quantile DEA: Estimating qDEA-alpha Efficiency Estimates with Conventional Linear Programming

Joseph A. Atwood and Saleem Shaik

Abstract Conventional non-parametric linear programming (LP) based data envelopment analysis (DEA) models have the advantage of being able to estimate multiple input-output efficiency metrics but suffer from sensitivity to outliers and statistical observational noise. Previous observation-deleting approaches to the outlier/noise problem have been somewhat ad hoc usually requiring iterative LP and non-LP problem solving methods. We present the theory and methodology of quantile-DEA (qDEA), similar in concept to quantile-regression, which enables the analyst to directly use LP to obtain efficiency metrics while specifying that no more than ψ-percent of data points can lie external to the efficiency hull. Estimated qDEA-α frontiers encompassing proportion $\alpha = 1 - \psi$ of the data observations are contrasted to order-α frontier estimates. Quantile DEA is shown to be useful in addressing outliers in a study examining changes in relative state level agricultural efficiency measures over time.

Keywords Data envelopment analysis · Partial moments · Outliers · Statistical noise · Quantile DEA

JEL Classification: Q18, C33, Q24

J.A. Atwood (✉)
Department of Agricultural Economics and Economics, Montana State University, Bozeman, MT, USA
e-mail: jatwood@montana.edu

S. Shaik
Department of Agribusiness and Applied Economics, North Dakota State University, Fargo, ND, USA
e-mail: saleem.shaik@ndsu.edu

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_14

1 Introduction

Data Envelopment Analysis (DEA) models are based on the output, input and graph distance functions developed by Malmquist (1953) and Moorsteen (1961) in the consumer context and Shephard (1953) in the producer context. The DEA approach to the study of efficiency has had a history in agriculture sector starting with M.J. Farrell (1957) and Farrell and Fieldhouse (1962). An operational research (OR) DEA model using linear programming (LP) was introduced by Charnes et al. (1978). DEA was popularized in a more informative and easily applied way by Fare et al. (1994). DEA gained popularity due to its ability to require little prior information with respect to a functional form, handle multiple outputs-inputs with caveats (Shaik et al. 2012; Shaik 2013) and strong/weak disposability assumptions (Shaik 1998). Due in part to these attractive features, over the past decade there has been a surge in the number of DEA empirical efficiency applications in numerous fields including agriculture, banking, health, sports, industrial regulation, and others (Emrouznejad et al. 2008; Liu et al. 2013). The Free Disposability Hull (FDH) was introduced by Deprins et al. (1984) and maintains free disposability while relaxing convexity.

While FDH and DEA non-parametric hull fitting has its advantages, there are also well known disadvantages. A key disadvantage of both FDH and DEA is the sensitivity of the estimated hulls or frontiers to statistical noise and data outliers. Several authors in the last decade have developed more robust quantile-like approaches under FDH or similar assumptions to obtain efficiency estimators that are less sensitive to outliers and statistical noise (Cazals et al. 2002; Aragon et al. 2005; Daouia and Ruiz-Gazen 2006; Daouia 2005; Daouia and Simar 2007; Daouia et al. 2008; Simar 2003; Simar et al. 2012). Of these approaches, Aragon et al's and Daouia and Simar's, FDH-related order-α estimator is similar to the concepts presented below and will be discussed in more detail later in the paper.

While various decision making unit (dmu) or observation deleting procedures have been used in DEA applications, quantile procedures have yet to be presented for linear programming based (LP) DEA models that incorporate convexity and constant (CRS), variable (VRS), decreasing (DRS) or increasing (IRS) returns to scale. We present a quantile LP based procedure that we term quantile Data Envelopment Analysis (qDEA)[1] build on Atwood et al. (1988) implementation of Atwood's (1985) partial moment stochastic inequality in a portfolio optimization problem. The qDEA approach presented is somewhat but not substantially more computationally intensive than traditional dual DEA methods. We present procedures for identifying dmu-specific quantile efficiency metrics as well as quantile-based DEA frontiers (denoted qDEA-α frontiers) similar in concept to order-α frontiers in that the projection points are identified by specifying that α percent of the points remain interior to the hull estimated relative to a given dmu. Given DEA's additional

[1]The terminology qDEA denotes quantile DEA not to be confused with Kousmanen and Post (2002) Quadratic DEA which they denote QDEA.

assumptions with respect to returns to scale, qDEA-α frontiers will differ from and tend to be smoother than order-α frontiers as we demonstrate below.

In the following discussion, we first review the mathematical theory that enables the construction of LP models that can endogenously identify subsets of the model's constraint set that can be violated. We then incorporate these results into DEA models that endogenously and simultaneously identify a set of firm specific influential points allowed to become 'super-efficient' and lie outside a given firm's estimated qDEA hull. We illustrate this procedure and present results using a modified example from Cooper et al. (2006). We then contrast estimated qDEA-α frontiers to order-α frontiers using the examples discussed in Simar and Wilson (2011a). Section 5 presents an empirical application of qDEA that examines changes in Nebraska state level agricultural efficiency scores over time. We conclude with a discussion of future research efforts and applications with respect to qDEA.

2 The qDEA Model

The qDEA process is implemented using two LP stages. The first stage (qDEA Stage-I) identifies the subset of 'quantile super-efficient' or external data points. The second stage (qDEA-stage-II) consists of a traditional DEA model where the constraints identified in the qDEA Stage-I model are relaxed. The qDEA Stage-I LP model is developed by modifying the LP model presented in Atwood et al. (1988) implementation of Atwood's (1985) partial moment stochastic inequality in a portfolio optimization problem. We note that Atwood et al.'s procedures are broadly applicable with linear programing problems where the modeler wishes to allow up to a pre-specified percentage of endogenously identified constraints to be relaxed or violated. In this section, we review the partial moment stochastic inequality and its use in relaxed constraint LP problems.[2] We then present a brief review of the traditional DEA model and the modifications required to implement the qDEA model.

2.1 Partial Moment Stochastic Inequalities

A lower partial moment (LPM) can be defined as:

$$\rho(\gamma, t) = \int_{-\infty}^{t} (t - x)^{\gamma} dF(x) \text{ for any } \gamma > 0. \tag{1}$$

[2]We wish to emphasize that while the qDEA process utilizes the results from a partial moment stochastic inequality to endogenously identify a set of external or "superefficient" DMUs, the set of external points are not randomly selected. This issue is discussed in more depth later in the paper.

Mean-LPM frontiers and tradeoffs have long been discussed in the finance literature as possibly attractive alternatives to Markowitz's mean-variance model. Fishburn (1972) presented several properties of LPMs including the relationship between mean-LPM efficient solutions and differing degrees of stochastic dominance. Semi-variance is a special case of an LPM ($\rho(2,\mu)$). In 1982, Berck and Hihn (BH) presented a stochastic inequality using semi-variance and demonstrated that their semi-variance based inequality often gave less conservative probability bounds than the one sided Chebychev inequality.

Atwood (1985) generalized Berck and Hihn's semi-variance stochastic inequality to the more general case using the LPM. The LPM stochastic inequality is derived as:

$$\begin{aligned}
\rho(\gamma,t) &= \int_{-\infty}^{t}(t-x)^{\gamma}dF(x)\\
&= \int_{-\infty}^{g}(t-x)^{\gamma}dF(x)+\int_{g}^{t}(t-x)^{\gamma}dF(x)\\
&\geq \int_{-\infty}^{g}(t-x)^{\gamma}dF(x)\\
&\geq \int_{-\infty}^{g}(t-g)^{\gamma}dF(x)=(t-g)^{\gamma}F(g)\\
&\Rightarrow \rho(\gamma,t)\geq(t-g)^{\gamma}F(g)\\
&\Rightarrow F(g)\leq\frac{\rho(\gamma,t)}{(t-g)^{\gamma}} \textit{for all } t>g,\gamma>0
\end{aligned} \tag{2}$$

The preceding results use Fishburn's lower partial moment but can easily be extended to the use of upper partial moments and computing limits on the probability of upside events.

The LPM-inequality is an interesting result that (with an appropriate choice of γ and t) will usually generate less conservative upper bounds than the one-sided Chebychev inequality $P(\mu-g\leq k\sigma)\leq\left(\frac{1}{1+k^2}\right)$ or the BH semi-variance inequality.

The linear partial moment ($\gamma=1$) inequality can often generate less conservative bounds than using higher order moments and we limit our discussion to the set of linear partial moments as linear partial moments can be computed in an LP model modeled with a finitely discrete set of outcomes. Atwood (1985) and Atwood et al. (1988) demonstrated that a linear partial moment's least restrictive level for t can be endogenously determined in a continuous linear programming model. Denoting the linear lower partial moment as $\rho(1,t)=\rho(t)$, Atwood showed that enforcing the constraints $t-\frac{1}{\psi}\rho(t)\geq g$ *and* $\rho\geq 0$ in a linear programming model is sufficient to guarantee that $F(g)\leq\psi$ for any positive ψ. To see this note that if t = g, $\rho(t=g)=0\Rightarrow F(g)=0<\psi$. If t > g rearranging the constraint gives $\frac{\rho(t)}{(t-g)}\leq\psi$ which combined with (2) implies that $F(g)\leq\psi$.

Atwood et al. constructed a model that maximized the expected income of a portfolio of assets subject to a set of technical constraints and the additional requirement that the proportion of discrete aggregate income outcomes falling below a target level g not exceed ψ. Modifying their notation slightly, Atwood et al.'s portfolio optimization model can be written as:

$$\begin{aligned} &\underset{x,t,d,\rho}{\text{Max}}\, \mu_Y x \\ &st. \\ &Ax \qquad\qquad\quad \leq b \\ &Yx - 1\mathrm{t} + Id \;\; \geq 0 \\ &\left(1/N\right) d - \rho \leq 0 \\ &\mathrm{t} - \tfrac{1}{\psi}\rho \qquad\quad \geq g \\ &x \geq 0, \mathrm{t} \geq 0, d \geq 0 \text{ and } \rho \geq 0 \end{aligned} \tag{3}$$

where μ is a vector of mean revenues, x is a vector of activity or portfolio weights, matrix A and vector b are technical portfolio constraint coefficients, Y is an (N by k) matrix of potential per unit income amounts, 1 is a vector of "ones", t is an endogenously determined upper limit on the linear LPM, d is a vector of deviations with $d_i > 0$ when $y_i x < \mathrm{t}$, $\left(1/N\right)$ is an vector of length N with elements equal to 1/N, ρ is the linear LPM, and g is a target level of income. System (3) selects a set of portfolio weights x that maximize the portfolio's expected income subject to technical portfolio constraints and the requirement that the number of aggregate income observations in the vector Yx falling below level g is less than ψN. The model effectively searches for a portfolio vector x that maximizes expected income subject to technical constraints while guaranteeing that no more than ψN of the N $Yx \geq g$ constraints are violated.

We conclude this discussion by noting that the portfolio chosen in system (3) will tend to be conservative in that the actual number of income observations falling below level g will often be less than the specified limit ψN due to the conservative nature of the LPM inequality.[3] In the following we utilize a two-stage process where a partial moment system similar to system (3) is utilized to identify constraints to be relaxed or eliminated. The second stage of the qDEA process consists of a conventional DEA model with the constraints identified in the first qDEA stage either being relaxed or deleted.

[3]The resulting portfolios were found in many cases to be conservative due to the use of the LPM inequality but the resulting solutions did satisfy the requirement that no more than ψN of the income constraints were violated. Subsequent unpublished research suggested that a two stage process similar to the two-stage qDEA process discussed below would generate less conservative outcomes while still satisfying the desired limit on the number of income observations falling below g.

2.2 Conventional DEA and Implementing Quantile DEA

Assuming there are K_X inputs, K_Y outputs and constant returns to scale (CRS), the conventional input orientation primal and dual LP's with outputs $k = 1, \ldots, K_Y$; inputs $i = 1, \ldots, K_X$; and DMU's $j = 1, .., N$ can be written as:

$$\begin{array}{ll} \textit{PRIMAL} & \textit{DUAL} \\ \text{Min}\left[0\ 1\right]\left[z\ \phi\right]' & \text{Max}\left[y_j\ 0\right]\left[p\ w\right]' \\ st.\begin{bmatrix} Y' & 0 \\ -X'\ x_j \end{bmatrix}\begin{bmatrix} z \\ \phi \end{bmatrix} \geq \begin{bmatrix} y_j \\ 0 \end{bmatrix} & st\begin{bmatrix} Y & -X \\ 0'\ & x_j' \end{bmatrix}\begin{bmatrix} p \\ w \end{bmatrix} \leq \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ z, \phi \geq 0 & p, w \geq 0 \end{array} \tag{4}$$

where $Y[j,k]$ is an $N x K_Y$ matrix of observed outputs; $X[j,i]$ is an $N x K_X$ matrix containing observed inputs, y_j is a vector containing the observed outputs for given DMU j, x_j is a vector containing the observed inputs for a given DMU_j, z is a vector of weights, ϕ is the efficiency measures of DMUj, p is a vector of output "prices" and w is the corresponding vector of input "prices".

The primal model searches for z (a set of projection weights) and estimate ϕ (the efficiency measure) of the given DMU. The dual model searches for vectors of prices (p, w) that maximize the efficiency score of the given DMU while requiring that the corresponding efficiency scores of all $DMU's$ be less than or equal to one. The effect of the DEA model can be viewed as fitting a hull around the data points, requiring all points to lie within the hull, and then constructing a distance metric measuring the proportional distance from a given $DMU's$ input-output combination to a point on the hull. With DEA, if the coefficients of input matrix X or output matrix Y contain statistical or outlier *noise*, the resulting hull and efficiency metric can be influenced by noise in the given $DMU's$ observations as well as noise in points near the efficiency frontier.

In the following, we present procedures for more robust qDEA efficiency estimates that allow the estimation of efficiency metrics while allowing no more than a given proportion ψ of the data points to lie external to the hull (or equivalently requiring at least proportion $\alpha = 1 - \psi$ of the points remain inside the fitted hull). We initially focus on the dual model as the dual's constraints more closely match the structure of the constraints in system (3).

2.3 Continuous Linear Programming qDEA Using Partial Moments

The traditional dual DEA model with N DMU's can be modified to implement the qDEA LP model. The input oriented DUAL qDEA model constructed by modifying expressions (3) and (4) can be written as:

$$
\begin{aligned}
&\textit{qDEA DUAL STAGE I}\\
&\text{Max}\left[\, y_j \;0\;0\;0\;0 \,\right]\left[\, p\;w\;t\;d\;\rho \,\right]'\\
&st\begin{bmatrix} Y & -X & 1 & -I & 0 \\ 0 & x_j & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{N} & -1 \\ 0 & 0 & -1 & 0 & \frac{1}{\psi} \end{bmatrix}\begin{bmatrix} p \\ w \\ t \\ d \\ \rho \end{bmatrix} \begin{matrix} \leq \\ \leq \\ \leq \\ \leq \end{matrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}\\
&p, w, T, d, \psi \geq 0
\end{aligned}
\tag{5}
$$

where[4] p is a vector of output prices, w is a vector of input prices, 1 is a vector of ones, t is the endogenously determined Partial Moment (PM) "integral" limit, I is an identity matrix, d is a vector of deviations above t, $1/N$ is an n-vector with each value equal to $1/N$, ρ is the endogenously calculated linear PM, and $0 < \psi < 1$ is the maximal proportion of data points that are allowed to lie outside the DEA hull.[5]

System (5) may sometimes be conservative in that fewer than 100ψ percent of the points may lie outside the DEA hull but the solution will never have more than 100ψ percent of the points outside DEA hull. The objective value from expression (5) is not our final efficiency estimate as system (5)'s objective value is an endogenously determined compromise between the desired objective function and a PM weighted distance metric and will be a biased estimate of the desired objective value. Below we demonstrate that the "extrapolated point" (and the resulting efficiency metric) from system (5) will consist of an extrapolation using information from all "external" points plus the new quantile DEA "support points."[6] As a result, we use system (5) in a first stage that endogenously and simultaneously identifies a set of points allowed to lie external to DMU's qDEA reference set. System (4) is then re-estimated in a second stage II where the constraints associated with the "external" points identified in stage I are relaxed or eliminated.

[4]Figures 2, 3 and 4 present a numerical example of the dual model in expression (4), and the qDEA model in expression (5).

[5]Setting $\psi < 1/N$ will guaranty that no points will lie outside the hull i.e., obtain the conventional DEA results.

[6]We use the terminology "support points" to denote the points that define the hyperplane onto which the given DMU's input-output points are projected with the distance from the initial point to the hyperplane being the estimated efficiency score. By a given DMU's "reference set" we refer to the set of points remaining on the same side of the projection hyperplane as the given DMU.

3 Example qDEA Model

3.1 A One Input: One Output Example

We demonstrate the implementation of the partial moment qDEA procedure by using a modification of Cooper, Seiford, and Tone's (CST) constant returns to scale single input (employees) – single output (sales) eight DMU example (page 26). Figure 1 reproduces CST's figure 2.1 where the input-output mix for point B has been changed to a more extremal value of (X, Y) = (3,4).[7] The following example demonstrates the use of qDEA while allowing one or two points to lie external to the efficiency hull.

With the traditional DEA analysis and CRS, the red ray through point B denotes the efficient frontier. The efficiency of all other points are measured relative to the red ray which lies at a substantial distance from all points except B. When one point is allowed to lie outside the hull, the model endogenously leaves point B external

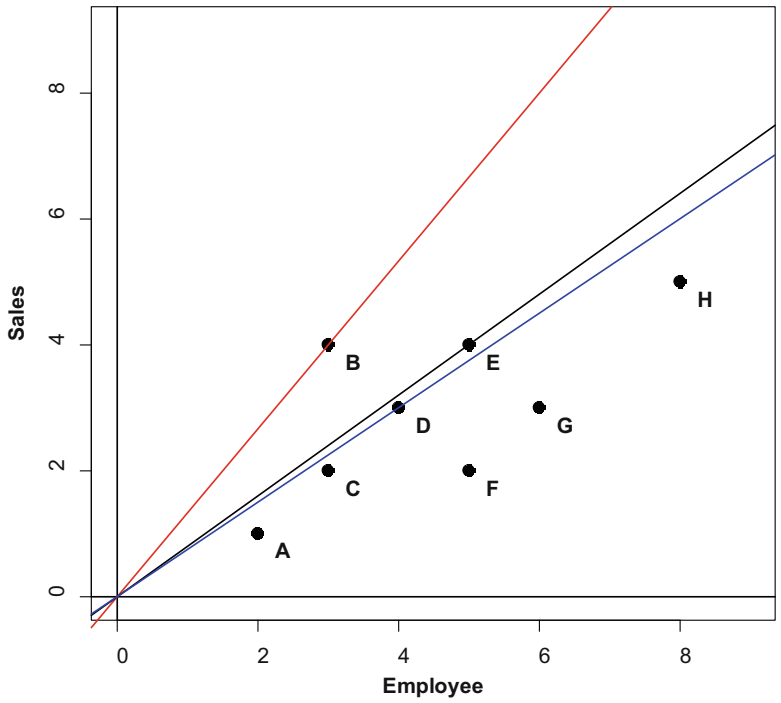


Fig. 1 Cooper, Seiford, and Tones single-input single output example with constant returns to scale

[7]Table 1 lists the input and out values used in this example.

CST EXAMPLE DATA			SELECTED			
DUAL PROBLEM			DMU =	H		
DMU	OUTPUT	INPUT		LHS	SIGN	RHS
A	1	-2		-0.1563	<=	0
B	4	-3		0.0000	<=	0
C	2	-3		-0.1875	<=	0
D	3	-4		-0.2188	<=	0
E	4	-5		-0.2500	<=	0
F	2	-5		-0.4375	<=	0
G	3	-6		-0.4688	<=	0
H	5	-8		-0.5313	<=	0
XRESTRICT	0	8		1	<=	1
OBJ	5	0		0.4688	<--OBJ	
	P	W				
VARS	0.09375	0.125				

Fig. 2 CST conventional dual DEA tableau DMU-H

to the hull with the resulting efficiency frontier changing to the black line through point E in Fig. 2. When two points are allowed to lie outside the hull, the model endogenously decides to leave points B and E outside the hull with the resulting efficiency hull changing to the blue line through point D.

The practical implementation of qDEA is illustrated in Figs. 2, 3 and 4 where we present screen shots from MS-Excel. Figure 2 presents an MS Solver screen shot of the conventional CRS DEA dual linear programming model and solution for DMU H. As expected, the solution indicates that point B is the restricting point on the DEA frontier giving a conventional DEA efficiency score of 0.4688 for point H.

The qDEA solution to this problem is obtained in two stages. Figure 3 presents an MS Solver tableau for first stage expression (5) of the qDEA procedure for a model allowing no more than two points to lie external to the hull. The first stage qDEA solution in Fig. 3 indicates that the model has determined to allow points B and E to lie external to the hull. From our experience with the conservative partial moment models, if we want no more than NP points to lie external to the hull, the solutions are somewhat less conservative if ψ is set just below (NP + 1)/N. For this problem with NP =2, ψ was set equal to (2 + 1)/8–0.0001 = 3/8–0.0001 = 0.3749 which will guarantee that the "probability" of the violating the constraints is 0.3749< 0.375 guaranteeing that that no more than two of the original constraints will be violated. The value $1/\psi$ in the constraint matrix is thus 1/0.3749 ∼2.6674. DMU H's objective value from qDEA Stage I is a conservative 0.6818 which we will discuss in more detail below.

CST EXAMPLE DATA		SELECTED		N out =	2	qDEA STAGE 1									
DUAL PROBLEM		DMU =	H	PLIM =	0.3749										
DMU	OUTPUT	INPUT	T	D-A	D-B	D-C	D-D	D-E	D-F	D-G	D-H	T-LPM	LHS	SIGN	RHS	DUALS
A	1	-2	1	-1	0	0	0	0	0	0	0	0	-0.02271	<=	0	0.00000
B	4	-3	1	0	-1	0	0	0	0	0	0	0	0.00000	<=	0	0.45464
C	2	-3	1	0	0	-1	0	0	0	0	0	0	-0.01136	<=	0	0.00000
D	3	-4	1	0	0	0	-1	0	0	0	0	0	0.00000	<=	0	0.45428
E	4	-5	1	0	0	0	0	-1	0	0	0	0	0.00000	<=	0	0.45464
F	2	-5	1	0	0	0	0	0	-1	0	0	0	-0.26136	<=	0	0.00000
G	3	-6	1	0	0	0	0	0	0	-1	0	0	-0.25000	<=	0	0.00000
H	5	-8	1	0	0	0	0	0	0	0	-1	0	-0.22729	<=	0	0.00000
XRESTRICT	0	8	0	0	0	0	0	0	0	0	0	0	1.00000	<=	1	0.68179
LPM CALC	0	0	0	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	-1	0.00000	<=	0	3.63716
QRESTRICT	0	0	-1	0	0	0	0	0	0	0	0	2.6674	0.00000	<=	0	1.36357
OBJ	5	0	0	0	0	0	0	0	0	0	0	0	0.681785	<--OBJ		
													NOTE THIS OBJ IS NOT THE FINAL ANSWER			
	P	W	T	D-A	D-B	D-C	D-D	D-E	D-F	D-G	D-H	T-LPM	STAGE 2 PROVIDES THE FINAL ANSWER			
VARS	0.136357	0.125	0.0909	0	0.2614	0	0	0.0114	0	0	0	0.0341				
# POINTS OUT HULL =	2		0	1	0	0	1	0	0	0						

Fig. 3 qDEA stage I tableau for model identifying two external points

CST EXAMPLE DATA			SELECTED	qDEA STAGE II		
DUAL PROBLEM			DMU =	H		
DMU	OUTPUT	INPUT		LHS	SIGN	RHS
A	1	-2		-0.08333	<=	0
B	4	-3		0.29167	<=	1000
C	2	-3		-0.04167	<=	0
D	3	-4		0.00000	<=	0
E	4	-5		0.04167	<=	1000
F	2	-5		-0.29167	<=	0
G	3	-6		-0.25000	<=	0
H	5	-8		-0.16667	<=	0
XRESTRICT	0	8		1	<=	1
OBJ	5	0		0.833333	<--OBJ	
	P	W				
VARS	0.166667	0.125				

Fig. 4 qDEA stage II tableau identifying qDEA efficiency scores

The second stage of the qDEA model is a conventional DEA model with relaxed constraints. Figure 4 presents the second stage tableau where the constraints associated with points B and E have been relaxed. The resulting efficiency metric is 0.8333 for point H. Note from Fig. 4 that we that relax the constraints associated with DMU's B and E by adding large positive values in the corresponding right hand side locations. From the primal perspective, this is equivalent to a "Big_M" method where we penalize the primal variables associated with DMU's B and E

Table 1 CST input orientation example – input-output combinations, CRS and VRS DEA and qDEA efficiency scores for one and two external points

					CRS		
				One External Point		Two External Points	
DMU	X	Y	DEA	qDEA-I(7)	qDEA-II(7)	qDEA-I(6)	qDEA-II(6)
A	2	1	0.375	0.500	0.625	0.545	0.667
B	3	4	1.000	1.333	1.667	1.455	1.778
C	3	2	0.500	0.667	0.833	0.727	0.889
D	4	3	0.563	0.750	0.938	0.818	1.000
E	5	4	0.600	0.800	1.000	0.873	1.067
F	5	2	0.300	0.400	0.500	0.436	0.533
G	6	3	0.375	0.500	0.625	0.545	0.667
H	8	5	0.469	0.625	0.781	0.682	0.833
					VRS		
				One External Point		Two External Points	
DMU	X	Y	DEA	qDEA-I(7)	qDEA-II(7)	qDEA-I(6)	qDEA-II(6)
A	2	1	1.000	1.250	1.500	1.333	1.500
B	3	4	1.000	1.333	1.667	1.556	2.000
C	3	2	0.778	0.833	1.000	0.889	1.000
D	4	3	0.667	0.750	1.000	0.833	1.000
E	5	4	0.600	0.800	1.000	0.933	1.200
F	5	2	0.467	0.500	0.600	0.533	0.600
G	6	3	0.444	0.500	0.667	0.556	0.667
H	8	5	1.000	NA	NA	NA	NA

severely enough that the LP model attempts to remove them from the basis. While the illustration in Fig. 4 is a dual model, qDEA Stage-II could be implemented in a more numerically efficient primal model with the appropriate variables removed or penalized via the Big M method.

3.2 qDEA Efficiency Scores for the CST Example

Table 1 presents the conventional DEA and the qDEA efficiency scores for all eight CST DMU's under both constant returns to scale (CRS) and variable returns to scale (VRS).

In Table 1 and following discussion, the notation qDEA-I(j) and qDEA-II(j) denotes the efficiency scores in the qDEA stage I or II model. Values of j less than one indicate quantiles or $\alpha = 1\text{-}\psi$ while j values exceeding 1 will denote the minimal number of points required to lie internal to the hull. Table 1 presents the output and input levels along with each DMU's estimated DEA, qDEA-I(j) and qDEA-II(j) efficiency scores with $j = 7$ and $j = 6$ points required to remain internal to the

data envelopment hull. We will also use the notation qDEA-α to denote the qDEA frontier when α percent of the points are internal or on the qDEA frontier (implying proportion $\psi = 1 - \alpha$ are allowed to be external to the qDEA frontier).

The Table 1 "DEA" values are the conventional input efficiency scores indicating that only DMU B is on the CRS efficient boundary (the red ray in Fig. 2) and DMU's A, B, and H are on the VRS efficient boundary (see Fig. 7 below). When one point is allowed to lie outside the hull, the CRS qDEA-II (7) efficiency scores indicate that the original point B is now "super-efficient" with a score greater than one (1.667) while point E becomes qDEA-II (7) efficient (1.00). The CRS qDEA-II (7) scores of all other points have increased by a substantial amount when contrasted to the original DEA efficiency estimates. The super-efficient score of 1.667 and the substantial increase of the other DMU's qDEA-II (7) efficiency scores provide evidence that point B is potentially an influential outlying point.

When two points are allowed external to the hull, the CRS qDEA-II (6) scores of all DMU's increase but by a smaller amount than when the first point was excluded. With two external points, DMUs B and E become qDEA super-efficient while DMU D is now on the efficient frontier.

Under variable returns to scale, DMU's A, B, and H are VRS DEA efficient with efficiency scores of 1. When 1 point is allowed to lie external to the hull, points A and B are super-efficient or lie external to the qDEA-(7/8) hull while DMU's C, D, and E lie on the qDEA-(7/8) hull. The NA's for DMU H indicate that the DMU H is super-efficient relative to the qDEA-(7/8) hull but the VRS qDEA LP problem is dual unbounded, primal infeasible for DMU H. The input-oriented primal infeasibility results because no movement the "input" direction from point H can reach the qDEA-(7/8) hull. As discussed below, with the VRS model when up to two points are allowed to lie external to the qDEA hull, the efficiency scores for DMU's A, C, D, F, and G do not change as it is not possible to find two points that can feasibly lie external to the hull given the DMU's in this example. The efficiency scores for DMU's B, and E can be increased by allowing points B and E to become simultaneously supper efficient.

3.3 Conservativeness of qDEA-I Efficiency Scores

As previously discussed, the qDEA-I efficiency scores are conservative relative to the qDEA-II efficiency scores. Figure 5 provides insights into the nature of the conservative solutions and why the two-stage process is used. Figure 5 plots the original points and the CRS DEA (red) and qDEA-II(6) (blue) efficient hulls from Fig. 1. The red and blue stars respectively plot the extrapolated efficient output-input combinations for DMU H for the DEA and qDEA-II(6) models. As expected, the red starred combination is extrapolated from point B while the blue starred efficient combination is extrapolated from point D. The green starred point is the output-input combination extrapolated by the first stage qDEA-I (6) model. The green

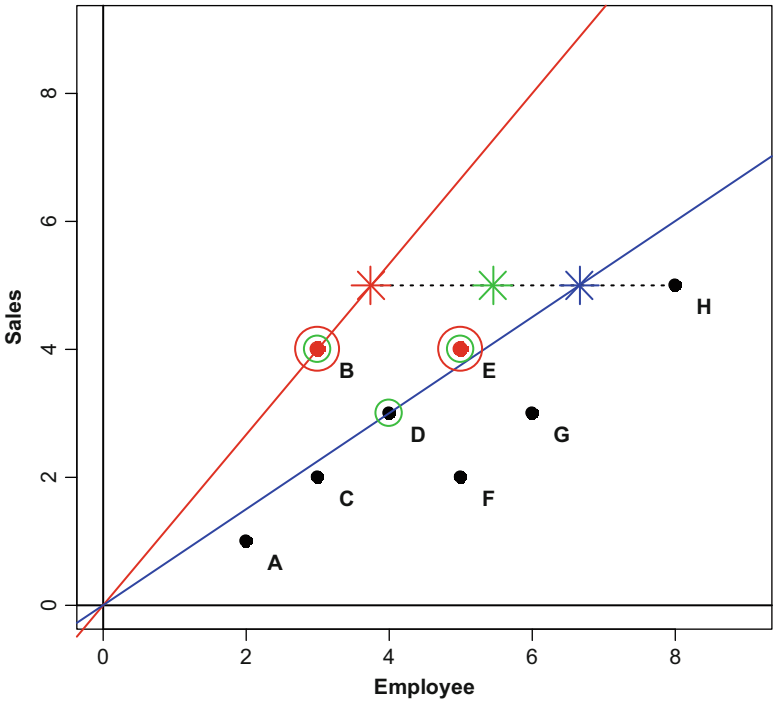


Fig. 5 CRS CST example DEA, qDEA-I, and qDEA-II extrapolated and excluded points

circled points (B, D, and E) are the points used in the qDEA-I (6) extrapolation with the red circled points denoting the qDEA-I (6) external points (positive d_i values in system (5)).[8] The efficiency score obtained using the green starred extrapolated point is clearly more conservative that the value associated with the blue starred extrapolation point.

We have presented the qDEA model and demonstrated its ability to obtain quantile efficiency metrics as well as qDEA's ability to identify sets of potentially outlying data points. We now discuss procedures for identifying qDEA-α frontiers and contrast the results to the conceptually similar order-α frontiers obtained using FDH related procdures.

[8]The fact that the green point is projected using each of the green circled points is readily derived by examining the dual solution to system (5) where the resulting dual or z_j values are projection weights. The dual values for all constraining equations in system (5) will be non-zero valued. The constraints associated with all external points as well the new qDEA support points will be binding in system (5).

4 Contrasting qDEA-α and Order-α Frontier Estimates

A qDEA-α "frontier" with respect to a given orientation **may** be defined as the set of all possible projection points that could be obtained by applying the qDEA process to points generated with a given data generating process while requiring that no more than $\psi = 1 - \alpha$ % of the points be allowed to lie external to any given DMU's qDEA reference set. This can be illustrated by examine the previous CST example under the assumption of variable returns to scale (VRS). Figure 6 plots the convex hulls for the sets of qDEA-1, qDEA-7/8, and the qDEA-6/8 projection points. The sets of projection points were obtained by solving the VRS qDEA model for each of the eight DMU's and computing the resulting projection points.

Figure 6 presents the qDEA-1, qDEA-7/8 and qDEA-6/8 frontiers respectively with black, red, and blue lines. Figure 6 also plots the projected points extrapolated from each DMU's original position. With the base DEA or qDEA-1 case, the VRS hull is the convex hull defined by points A-B-H. The frontier defined by these points is plotted in black. The "stars" on the black line denote points projected from

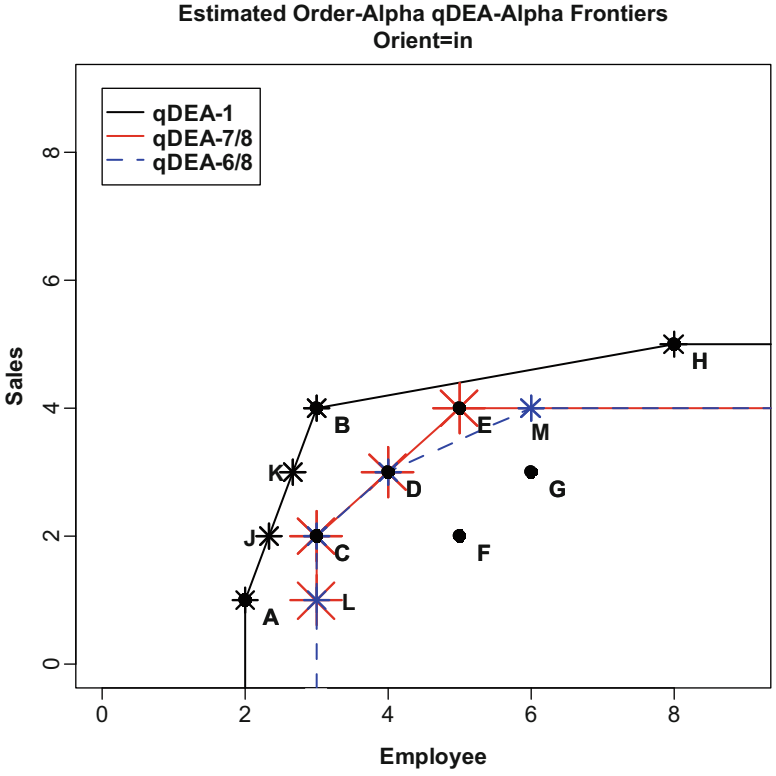


Fig. 6 VRS CST example DEA, qDEA-I, and qDEA-II extrapolated points and qDEA-α frontiers

the DMU points with DMU's A, B, and H being VRS-efficient, points C and F projecting to point J, points D and G projecting to point K, and point E projecting to point B.

With up to one external point, the efficiency scores of all DMU's (with the exception of DMU H) improves as each of DMU's A-G's input-output points are now projected onto the red qDEA-7/8 frontier. Point A now projects to point L, point B now projects onto point E, points C, D, and E are on the qDEA-7/8 frontier, point F projects to point C and point G now projects to point D. As indicated previously, point H's dual solution is unbounded indicating an infeasible primal solution as no horizontal shift from point H can reach the qDEA-7/8 red frontier. In this case, we note that DMU H is qDEA-7/8 "super-efficient" but cannot recover a qDEA-7/8 distance for DMU H. DMU's A and B are identified as qDEA-7/8 "supper-efficient in that their efficiency scores exceed 1 and in each case the bounded solution indicates that each DMU's input-output values can be projected back onto the qDEA-7/8 frontier.

With VRS and up to two external points, we find that only DMU B and E's efficiency scores increase beyond the values realized with one external point. The qDEA-6/8 efficiency scores for the remaining DMU's remain unchanged as it is not possible to find only two points that can lie external to each point's reference set. Consider DMU A's possibilities. In the qDEA-7/8 model, DMU A's efficiency score was improved by allowing point A to lie external to the set and projecting point A to point L on the qDEA-7/8 frontier. With up to two external points and VRS, we find that point A could only be projected further to the right of point L if point A and both points B and C were allowed to lie external to the qDEA-6/8 frontier. This would require allowing up to three external points rather than two. As a result, point L remains on both the qDEA-7/8 and the qDEA-6/8 frontiers. Similar arguments apply to points C and D. However, for DMU's B and E, we find that their qDEA-6/8 efficiency scores can be improved by allowing both points B and E to be super-efficient with their resulting local qDEA-6/8 frontier projected point M being constructed as a convex combination of support points D and H. The net result of this exercise is that points A, B, and H define the VRS DEA and qDEA-1 frontier, points L, C, D, and E define the VRS qDEA-7/8 frontier, and points L, C, D, and M define the qDEA-6/8 frontier. Using this methodology, we can construct qDEA-α frontiers that can be contrasted to order-α frontiers illustrating some of the differences between the two types of frontiers.

4.1 qDEA-α and Order-α Frontiers

To facilitate our comparison of qDEA-α and order-α frontiers we utilize the example presented in Simar and Wilson (2011a). Simar and Wilson's example involves one input and one output with inputs distributed over the [0,1] interval and realized production in the domain $0 \leq y \leq \sqrt{(2x - x^2)}$. Joint realizations of x and y are assumed to be uniformly distributed over the resulting x-y space with density

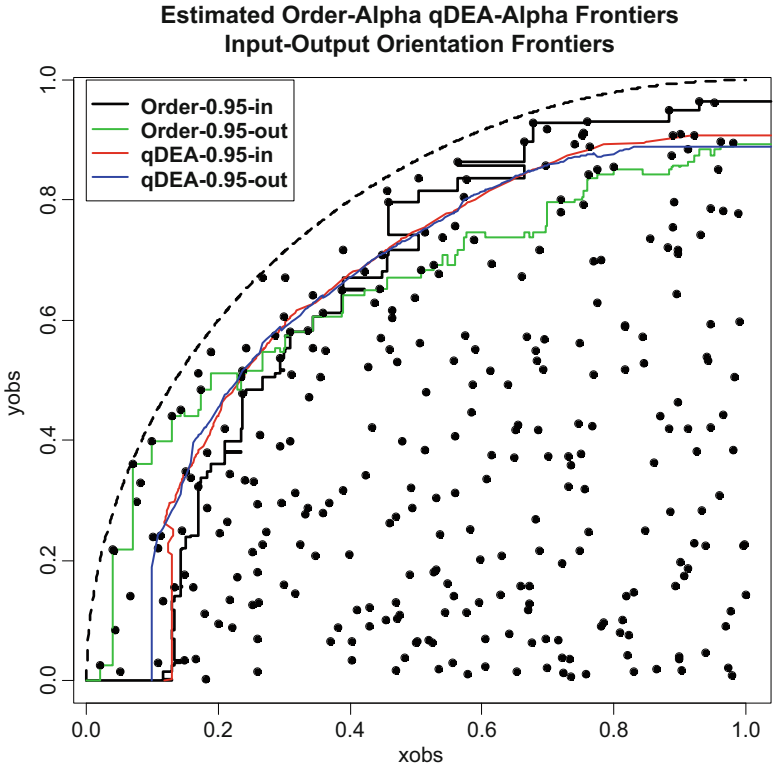


Fig. 7 Estimated order-α and qDEA-α frontiers with 300 DMU's

function $f(x, y) = 4/\pi$ for $x \in [0, 1]$ and $0 \leq y \leq \sqrt{(2x - x^2)}$. In the following comparisons, random joint realizations of x and y were generated using this process and the resulting qDEA-α and order-α frontiers calculated using the "frontiles" R package (Daouia and Laurent (2013)) and qDEA R code that is available from the authors upon request.[9] Both the "frontiles" package and the author's R code can generate input and output orientation frontiers. Fig. 7 presents a combined set of VRS input and output qDEA-0.95 and order-0.95 estimated frontiers using data from 300 simulated DMU observations generated with the above process.

In Fig. 7 the dashed black line represents the upper boundary of $0 \leq y \leq \sqrt{(2x - x^2)}$, the black line is the estimated input-oriented order-0.95 frontier, the green line is the estimated output-oriented order-0.95 frontier, the red line is the estimated input-oriented qDEA-0.95 frontier, and the blue line is the estimated output-oriented qDEA-0.95 frontier. The reader will note that the order-α frontiers

[9]R code available from the authors can be used to obtain qDEA solutions with input, output, or the more general ddea DEA models.

are consistent with the examples presented by Simar and Wilson with the upper right end of the black input order-0.95 and the lower left ends of green output order-0.95 frontiers approaching the dashed upper production boundary. As discussed by Simar and Wilson the order-α boundaries tend to be somewhat jagged with a sample size of 300 DMU's. The reader will note that the qDEA-0.95 frontiers input and output oriented frontiers coincide closely with their order-0.95 input and output counterparts respectively in the lower left and upper right sections of the frontiers. However, the qDEA-0.95 frontiers are substantially smoother than their order-0.95 counterparts over most of the frontiers' range and the input and output qDEA-0.95 frontiers align much more closely with each other over most of the frontiers' ranges as well. The smoother VRS DEA based qDEA-0.95 frontiers are to be expected given the structure imposed by the VRS assumptions in comparison to the more general FDH-related methodologies used in estimating the order-α frontiers.

5 An Application of qDEA in a Study Examining Changes in State Level Agricultural Efficiencies Over Time

The potential usefulness of qDEA is illustrated in a study examining changes in U.S state level agricultural efficiencies over time. As space precludes a discussion of the results for all states, we focus upon estimated changes in the state of Nebraska's estimated efficiencies over the period 1964 to 2014.

State level aggregate agricultural input-output data for the period 1960 to 2004 were obtained from ERS (2016). The data consists of several variables including standardized values (1996$) of total agricultural output, livestock output, crop output, capital input services (excluding land), land input service flows, labor input services, and intermediate inputs including items such as energy, chemicals, and fertilizer. For the purposes of this example, we utilize aggregate agricultural output as our single output variable. Two inputs, "fixed" and "variable" inputs were constructed. "Fixed" inputs were constructed as the sum of the non-land capital and land service values. "Variable" inputs were constructed as the sum of the labor and total intermediate input values.

Agricultural input and output data is inherently volatile due to the effects of weather. To reduce the effects of weather-induced variability, we computed a moving 5-year Olympic average for each of the output and two input variables leaving a data set with observations from 1964–2004. Due to heterogeneity in the scale of agriculture across states, we then divided each state's input and output levels by the year's state level output for each year in the 1964–2004 period. Each state's output is thus 1 and the two inputs are standardized to a per-unit-of-output basis for each year in the 1964–2004 series.

Figure 8 presents standardized Fixed-Variable input plots for the years 1974, 1984, 1994, and 2004 with Nebraska's observations plotted in red. Several observations can be made from a visual examination of the plots. We first note that most states' observed absolute efficiencies have improved over time with fewer of

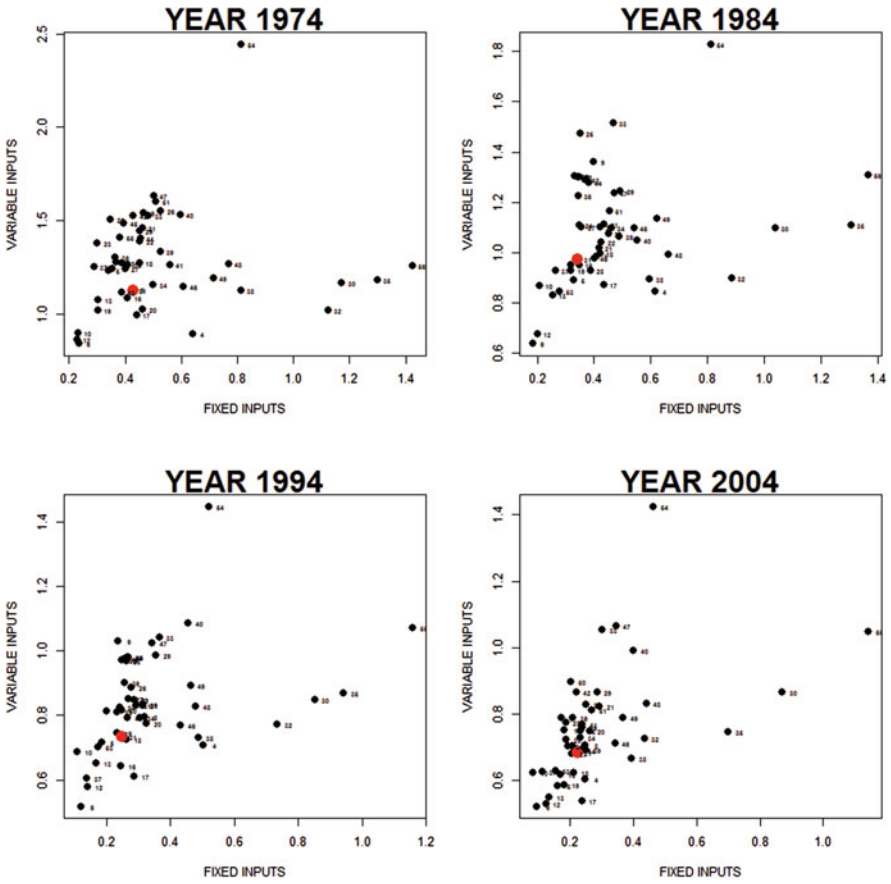


Fig. 8 Fixed – variable inputs per unit of output 1974–2004

both variable and fixed inputs being required to produce a unit level of output. We also note that the data has persistent outliers possibly generated by differences in the structure and types of agriculture across states. In all 4 years, a group of two or more isolated observations (including California and Florida) are observed in the lower left portion of the plots with other isolated points located in the right and upper portions of the plots. Assuming constant returns to scale (CRS), the points in the lower left portion of the plot will likely be "influential" outliers that may substantially affect the estimated efficiency scores of the remaining states' observations. While the isolated points in the right and top portions of the plot are "outlier's" with respect to most of the other states, they will likely be "non-influential" outliers that have no effect upon the other states' estimated efficiency scores.

Given differences in the types of agriculture, it is questionable whether California and Florida should be included in the group of peer states to which Nebraska is

compared. While California and Florida could have been excluded from the data set, there appear to be additional "lower left" states in the later years and we did not want to exogenously exclude one or more states from the analysis in an ad. hoc. Manner. The qDEA procedure allow us to systematically examine the effects of using alternative "quantile-based" peer groups by endogenously identifying sets of external points and computing efficiency scores relative to the remaining peer groups. To examine changes in a state's efficiency scores over time, a time series of input-orientation efficiency scores were estimated for each state by looping through the years 1964–2004, subsetting the data for each year, and estimating both traditional and quantile efficiency scores. For each state, the estimated efficiency scores were then plotted against time.

Figure 9 plots the time varying efficiency scores for the state of Nebraska with from one to four external or quantile super-efficient points. The traditional

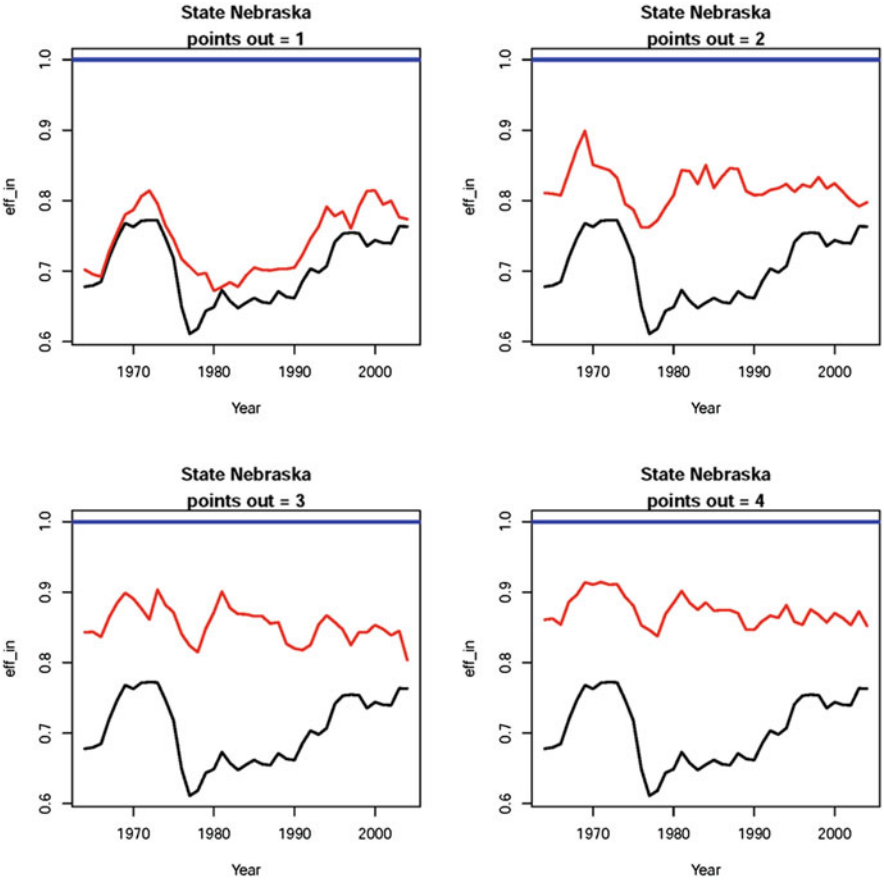


Fig. 9 Nebraska's estimated relative agricultural efficiencies over time and by number of external qDEA observations

DEA results (the black lines), would imply that Nebraska's estimated efficiencies initially rose sharply, dropped sharply in the mid 70's and rose steadily from the late seventies through 2004. To examine the sensitivity of these results to the possible presence of influential outliers, we estimated a series of qDEA efficiency scores. Nebraska's intertemporal qDEA efficiency are plotted in red in Fig. 9. With one external point Nebraska's estimated qDEA efficiency scores increase but the exhibited pattern is similar to the original DEA results with an initial increase and decline in efficiency scores and then a steady increase from the late 1970s through 2004.

However, with two or more external points, Nebraska's intertemporal efficiency score patterns stabilize to a pattern with fluctuations but little or no trend in efficiency scores over time. Further increasing the number of external points from 4 to 10 increased the absolute levels of the efficiency scores but resulted in no substantive change in the intertemporal pattern i.e. Nebraska's relative efficiency scores exhibited little or no change over the period 1964–2004.

6 Summary and Conclusions

This paper has presented procedures that allow the analyst to implement Quantile Data Envelopment Analysis where efficiency metrics are developed while allowing no more than a specified proportion of the data observations to lie outside the estimated efficiency hull. The procedures are straight forward and can be easily implemented with conventional linear programming algorithms. An additional potential use of qDEA is in more easily identifying potential comparative peer or benchmarking groups within a set of possibly technically heterogeneous firms. Benchmarking studies often involve contrasting a firm's characteristics and performance against metrics from a group of industry firms such as the top 10 percent or top quartile of firms. The ability to identify quantile DEA efficiency metrics allows the analyst to now "efficiently" perform similar comparisons using DEA efficiency estimates and distance metrics.

The qDEA approach is not without its limitations. The researcher will need to experiment with the effects of using differing quantiles upon the estimated efficiency metrics in specific applications. The added complications of appropriate quantile selection are similar to selecting the appropriate quantile level in quantile regressions. The qDEA procedure is still in its infancy and will benefit from further development. Additional research is needed with respect to the statistical properties of the qDEA efficiency estimates,[10] the selection of appropriate quantile levels, the

[10]The authors, to date, have not been able to derive closed form expressions for the asymptotic properties of the qDEA estimates. Numerical procedures described by Geyer (2013), suggest that qDEA estimates appear to have many of the desirable features of the FDH related order-m and order-α estimators including root-n convergence and asymptotic normality. The authors have experimented extensively with the use of nCm bootstrapping as discussed by Politis et al. (1999,

use of qDEA to identify groups of potentially outlying points, and its usefulness in identifying quantile based peer groups. However, as clearly demonstrated with the application discussed above, the ability to systematically and endogenously examine the effects of allowing proportions of points to lie external to the efficiency hull should prove useful to the researcher when faced with statistical noise and potential outliers.

References

Aragon, Y., Daouia, A., & Thomas-Agnan, C. (2005). Nonparametric frontier estimation: A conditional quantile-based approach. *Econometric Theory, 21*(2), 358–389.

Atwood, J. A. (1985). Demonstration of the use of lower partial moments to improve safety-first probability limits. *American Journal of Agricultural Economics, 67*, 787–793.

Atwood, J. A., Watts, M. J., Helmers, G. A., & Held, L. J. (1988). Incorporating safety-first constraints in linear programming production models. *Western Journal of Agricultural Economics, 13*, 29–36.

Berck, P., & Hihn, J. M. (1982). Using the semivariance to estimate safety-first rules. *American Journal of Agricultural Economics, 64*, 298–300.

Cazals, C., Florens, J.-P., & Simar, L. (2002). Nonparametric frontier estimation: A robust approach. *Journal of Econometrics, 106*, 1–25.

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

Cooper, W., Seiford, L., & Tone, K. (2006). *Introduction to data envelopment analysis and its uses*. New York: Springer.

Daouia, A. (2005). Asymptotic representation theory for nonstandard conditional quantiles. *Journal of Nonparametric Statistics, 17*, 253–268.

Daouia, A., & Laurent, T. (2013). *Frontiles: Partial frontier efficiency analysis*. R package version 1.2. https://CRAN.R-project.org/package=frontiles

Daouia, A., & Ruiz-Gazen, A. (2006). Robust nonparametric frontier estimators: Influence function and qualitative robustness. *Statistica Sinica, 16*, 1233–1253.

Daouia, A., & Simar, L. (2007). Nonparametric efficiency analysis: A multivariate conditional quantile approach. *Journal of Econometrics, 140*, 375–400.

Daouia, A., Florens, J.-P., & Simar, L. (2008). *Frontier estimation and extreme values theory*, preprint-submitted paper for publication. Available at http: www.stat.ucl.ac.be

Deprins, D., Simar, L., & Tulkens, H. (1984). Measuring labor efficiency on post offices. In P. Marchand & H. Tulkens (Eds.), *The performance of public enterprises: Concepts and measurement* (pp. 243–267). Amsterdam: North Holland.

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

ERS (2016). Table 24. https://www.ers.usda.gov/data-products/agricultural-productivity-in-the-us/agricultural-productivity-in-the-us/#State-Level Tables, Price Indices and Implicit Quantities of Farm Outputs and Inputs by State, 1960–2004.

2001), Geyer (2013) and Simar and Wilson (2011b) and have achieved nominal qDEA confidence interval coverage levels that compare favorably with conventional DEA results reported by Simar and Wilson.

Fare, R., Grosskopf, S., & Lovell, C. A. K. (1994). *Production frontiers*. Cambridge: Cambridge University Press.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society Series A, 120*, 253–281.

Farrell, M. J., & Fieldhouse, M. (1962). Estimating efficient production functions under increasing returns to scale. *Journal of the Royal Statistical Society., 125*, 252–267.

Fishburn, P. J. (1972). Mean-risk analysis with risk associated with below-target returns. *American Economic Review., 67*, 116–126.

Geyer, C. J. (2013). *Statistics 5601 Notes: The Subsampling Bootstrap*. University of Minnesota. Reference Material : http://www.stat.umn.edu/geyer/5601/examp/subboot.html

Kousmanen, T., & Post, T. (2002). Quadratic data envelopment analysis. *Journal of the Operational Research Society, 53*, 1204–1214.

Liu, J. S., Louis, Y. Y. L., Wen-Min, L., & Lin, B. J. Y. (2013). Data envelopment analysis 1978–2010: A citation-based literature survey. *Omega, 41*(1), 3–15. https://doi.org/10.1016/j.omega.2010.12.006.

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

Moorsteen, R. H. (1961). On measuring productive potential and relative efficiency. *Quarterly Journal of Economics, 75*, 451–467.

Politis, D. N., Romano, J. P., & Wolf, M. (1999). *Subsampling*. New York: Springer.

Politis, D. N., Romano, J. P., & Wolf, M. (2001). On the asymptotic theory of subsampling. *Statistica Sinica, 11*, 1105–1124.

Shaik, S. (1998). *Environmentally Adjusted Productivity [EAP] Measures for Nebraska Agriculture Sector*. Ph.D. dissertation Department of Agricultural Economics, University of Nebraska-Lincoln.

Shaik, S. (2013). Does crop insurance affect the technical efficiency? a panel DEA analysis. *American Journal of Agricultural Economics., 95*(5), 1136–1154.

Shaik, S., Mishra, A., & Atwood, J. A. (2012). Aggregation issues in the estimation of linear programming productivity measures. *Journal of Applied Economics, XV*(1), 169–187.

Shephard, R. W. (1953). *Cost and production functions*. Princeton: Princeton University Press.

Simar, L. (2003). Detecting outliers in frontier models: a simple approach. *Journal of Productivity Analysis., 20*, 391–424.

Simar, L., & Wilson, P. W. (2011a). Estimation and inference in nonparametric frontier models: Recent developments and perspectives. *Foundations and Trends in Econometrics, 5*(3–4), 183–337.

Simar, L., & Wilson, P. W. (2011b). Inference by the m Out of n Bootstrap in nonparametric frontier models. *Journal of Productivity Analysis, 36*, 33–53.

Simar, L., Vanhems, A., & Wilson, P. W. (2012). Statistical inference for DEA estimators of directional distances. *European Journal of Operational Research, 220*, 853–864.

Erratum to: Estimating Efficiency in the Presence of Extreme Outliers: A Logistic-Half Normal Stochastic Frontier Model with Application to Highway Maintenance Costs in England

Alexander D. Stead, Phill Wheat, and William H. Greene

Erratum to:
Chapter 1 in: W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_1

The book was inadvertently published with an incorrect affiliation of the author W.H. Greene in Chapter 1 as "University of New York" whereas it should be "New York University".

The updated online version of this chapter can be found at https://doi.org/10.1007/978-3-319-68678-3_1

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3_15

Index

W.H. Greene et al. (eds.), *Productivity and Inequality*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-319-68678-3