Malmquist and Törnqvist Productivity Indexes: Returns to Scale and Technical Progress with Imperfect Competition

W. Erwin Diewert and Kevin J. Fox[1] February 15, 2010.
Discussion Paper 10-03,
Department of Economics,
University of British Columbia,
Vancouver, Canada, V6T 1Z1.

Email addresses: diewert@econ.ubc.ca and K.Fox@unsw.edu.au

Abstract

Caves, Christensen and Diewert proposed a method for estimating a theoretical productivity index for a firm using Törnqvist input and output indexes, augmented by exogenous estimates of local returns to scale. However, in order to implement their method, they assumed that the firm maximized revenue in each period, conditional on the observed input vector in each period, taking output prices as fixed. This assumption is not warranted when there are increasing returns to scale. Thus in the present paper, it is assumed that the firm solves a monopolistic profit maximization problem when there are increasing returns to scale and the results of Caves, Christensen and Diewert are modified in accordance with this assumption.

Key Words

Productivity, index numbers, Malmquist indexes, Törnqvist indexes, returns to scale, non-competitive behavior, flexible functional forms.

Journal of Economic Literature Classification Numbers

C43, D24, E23

[1] The authors are grateful for helpful comments from Bert Balk, Alice Nakamura and two referees, and for financial support from the Australian Research Council, the SSHRC of Canada, and the Ministry of Education and Science of Spain (Secretaría de Estado de Universidades e Investigación, SAB2003-0234), along with the hospitality of the University of Valencia. None of the above are responsible for opinions expressed in the paper

1. Introduction

The application of the Malmquist (1953) productivity index by Caves, Christensen and Diewert (1982) (CCD) to a flexible functional form in an exact index number context has found many applications in diverse contexts.[2] Applications include the assessment of the productive performance of countries[3], regulated utilities[4], agriculture in developing countries[5], financial institutions[6], dialysis markets[7] and polluting firms[8].

The CCD framework uses the "economic approach" to justifying the choice of index number formulae for calculating aggregate indexes of input, output and productivity. This approach specifies the index of interest in terms of a theoretical (Malmquist) index, assumes a particular functional form to represent the underlying technology, and then derives the index number formula which corresponds to the theoretical index. An index number formula derived in such a way provides a straightforward means of estimating the underlying theoretical index.

However, the CCD results have two weaknesses:

- Their results are derived under the assumptions that producers minimize costs taking input prices and output targets as fixed and that they also separately maximize revenue taking output prices as fixed and inputs as fixed;
- If there are increasing returns to scale, then exogenous estimates of the (local) degree of returns to scale are required in order to evaluate empirically their productivity measure.

The present paper shows how the above two problems can be overcome. The assumption of competitive revenue maximizing behavior will be replaced by the assumption of monopolistic profit maximization if there are increasing returns to scale. The paper will also show how a simple one equation econometric model consistent with the underlying theoretical framework can be derived that will enable researchers to estimate the degree of returns to scale. Thus the paper will extend the results of CCD in order to demonstrate how a standard Törnqvist productivity index, derived from a theoretical Malmquist index, can be decomposed into technical change and returns to scale components. The Törnqvist productivity index is used by various agencies around the world to measure productivity growth, for example, by the U.S. Bureau of Labor Statistics (2002). Although the Törnqvist index formula has the form of a weighted geometric mean which allows useful

[2] This theoretical productivity index was independently proposed by Hicks (1961) and Moorsteen (1961) and is based on the distance function idea originally introduced by Malmquist (1953) in the consumer context.

[3] See Färe, Grosskopf, Norris and Zhang (1994), Kumar and Russell (2002) and Kruger (2003).

[4] See Atkinson, Conwell and Honerkamp (2003) and Coelli, Estache, Perelman and Trujillo (2003).

[5] See Nin, Arndt and Preckel (2003).

[6] See Alam (2001) and Sturm and Williams (2004).

[7] See Ozgen and Ozcan (2004).

[8] See Hailu and Veeman (2000) and Weber and Domazlicky (2004). For a range of other applications and references, see e.g. Färe, Grosskopf and Russell (1998), Fox (2002), and Cooper, Seiford and Zhu (2004). For recent theoretical advances, see Färe and Grosskopf (2004), De Borger and Kerstens (2000) and Briec and Kerstens (2004). For a review of available software packages for the estimation of Malmquist productivity indexes, see Hollingsworth (2004).

decompositions in many contexts,[9] these decompositions do not extend naturally to productivity indexes when there are increasing returns to scale.

Conventional productivity growth, defined as an output growth index divided by an input growth index, can be driven by movements in the technology frontier (technical progress)[10] as well as movements along the frontier (returns to scale). The latter effect implies that returns to scale may be the cause of fluctuations in the index of conventional productivity growth. It is useful to theoretically and empirically determine the respective roles of technical progress and returns to scale. As well as the basic issue of gaining an understanding of the sources of productivity growth, the reasons for this interest include the recent increase in economic growth models with increasing returns to scale,[11] the implications for understanding the role of returns to scale in industrial organization[12] and the related implications for regulation.[13]

In order to identify the sources of change in a productivity index, one needs to consider the underlying economic justification for the index. CCD provide such a justification, but their assumptions about producer behavior are not entirely satisfactory. Our results here are achieved in the context of a conventional model of imperfect competition, meaning that the usual assumption of price taking behaviour which CCD and others used to establish relationships between underlying economic functions and index number formulae is in fact unnecessary. This means that the economic approach to index numbers can be used to justify the use of exact index number formulae for productivity assessment even in non-competitive environments, such as the case of firms in regulated industries. This greatly strengthens the theoretical underpinnings of empirical analysis in this context.

Finally, empirical implementation of the method for determining the role of returns to scale and technical progress yields statistical error terms, which can be interpreted as productivity "shocks" of the type of interest in many macroeconomic modelling contexts.

In section 2 below, we provide the basic theoretical definitions for the Malmquist input, output and productivity indexes, and in section 3, we provide a simple method for separating technical progress and returns to scale for a Törnqvist productivity index derived from a Malmquist index using the economic approach to index numbers. Section 4 concludes.

2. Malmquist Input, Output and Productivity Indexes

There has been considerable recent interest in, and debate concerning, alternative approaches to decomposing the Malmquist productivity index introduced by CCD; for a

[9] See Kohli (2003), Fox, Grafton, Kirkley and Squires (2003) and Shui (2003).

[10] See Tinbergen (1942), Solow (1957) and Jorgenson and Griliches (1967), who assumed constant returns to scale.

[11] See Bennett and Farmer (2000), Guo and Lansing (2002), Hintermaier (2003), Jones (2004), Guo (2004) and Benhabib and Wen (2004).

[12] See Ciccone (2002), Norman and Venables (2004) and Wang (2003).

[13] See McIntosh (2002).

review of the issues and the debate, see Balk (2001) and Grosskopf (2003). Here we give the basic theoretical definitions for the Malmquist input and output indexes and a preliminary definition for the Malmquist productivity index, following fairly closely the definition of these indexes by CCD.

Let S^t be the *production possibilities set* for a production unit or firm for periods t = 0,1. We assume that S^t is a nonempty closed subset of the nonnegative orthant in Euclidean M+N dimensional space. If (y,x) belongs to S^t, then the nonnegative vector of M outputs $y \equiv [y_1,...,y_M] \geq 0_M$ can be produced using the period t technology by the vector of N nonnegative inputs $x \equiv [x_1,...,x_N] \geq 0_N$.[14]

Using the period t production possibilities set S^t and given a strictly positive output vector $y >> 0_M$ and a strictly positive input vector $x >> 0_N$, the production unit's period t *input distance function* D^t for periods t = 0,1 can be defined as follows:

$$(1)\ D^t(y,x) \equiv \max_{\delta>0} \{\delta : (y,x/\delta)\in S^t\}.$$

Thus given the strictly positive vector of outputs y and the strictly positive vector of inputs x and the period t technology S^t, $D^t(y,x)$ is the maximal amount that the input vector x can be deflated so that the deflated input vector $x/D^t(y,x)$ can produce the vector of outputs y. Denote the period t observed production vector for the production unit by (y^t,x^t) for t = 0,1. If the period t production vector is on the frontier of the period t production possibilities set, then it can be seen that the period t input distance function, $D^t(y^t,x^t)$, is equal to one.

Instead of deflating the input vector x so that the resulting deflated vector is just big enough to produce the vector of outputs y, we could think of deflating the output vector so that the resulting deflated output vector is just producible by the input vector x. Thus given $y >> 0_M$ and $x >> 0_{N,}$ the production unit's period t *output distance function* d^t for periods t = 0,1 can be defined as follows:

$$(2)\ d^t(y,x) \equiv \min_{\delta>0} \{\delta : (y/\delta,x)\in S^t\}.$$

It is not immediately clear that the maximum in (1) or the minimum in (2) will exist. In fact, in order to obtain the existence of the functions D^t and d^t defined by (1) and (2), some restrictions on the production possibilities sets S^t are required (in addition to the assumption that S^t is a closed, nonempty subset of the nonnegative orthant). In the technical Appendix, we postulate a simple set of restrictions on the S^t which will guarantee the existence of these input and output distance functions.

CCD did not use definitions (1) and (2) in order to define the input and output distance functions. Instead, they defined D^t and d^t in an equivalent manner using the production unit's period t *production function*, F^t, or the firm's *input requirements function*, g^t. We will now explain how these functions can be defined, given the production possibilities sets, S^t.

[14] Notation: $y \geq 0_M$ means each component of the vector y is nonnegative, $y >> 0_M$ means that each component is strictly positive, $y > 0_M$ means $y \geq 0_M$ but $y \neq 0_M$ and $p{\cdot}y$ denotes the inner product of the vectors p and y.

Given S^t and nonnegative output and input vectors, y and x, we rewrite the output vector $y \equiv [y_1,y_2,...,y_M]$ as $[y_1,\tilde{y}]$ where $\tilde{y}$, the vector of "other than y_1 outputs", is defined as the vector $[y_2,...,y_M]$. The *period t production function*, F^t, is defined as follows:

(3) $F^t(\tilde{y},x) \equiv \max_{y_1} \{y_1 : (y_1,\tilde{y},x)\in S^t\}$; $t = 0,1.$

If there is no y_1 such that $(y_1,\tilde{y},x)\in S^t$, then we define $F^t(\tilde{y},x) = -\infty$. Basically, $F^t(\tilde{y},x)$ is the maximum amount of the first output that can be produced in period t by the production unit, given that it also produces the nonnegative vector of other outputs $\tilde{y}$ and given that it has the nonnegative vector of inputs x at its disposal.

Given S^t and nonnegative output and input vectors, y and x, we rewrite the input vector $x \equiv [x_1,x_2,...,x_N]$ as $[x_1,\tilde{x}]$ where $\tilde{x}$, the vector of "other than x_1 inputs", is defined as the vector $[x_2,...,x_N]$. The *period t input requirements function*, g^t, is defined as follows:

(4) $g^t(y,\tilde{x}) \equiv \min_{x_1} \{x_1 : (y,x_1,\tilde{x})\in S^t\}$; $t = 0,1.$

If there is no x_1 such that $(y,x_1,\tilde{x})\in S^t$, then we define $g^t(y,\tilde{x}) = +\infty$. Fundamentally, $g^t(y,\tilde{x})$ is the minimum amount of the first input that is required in period t in order to produce the vector of outputs y given that the production unit has the nonnegative vector of other inputs $\tilde{x}$ at its disposal.

As mentioned above, CCD used the functions F^t and g^t in order to develop their results. We will now outline some of their key definitions and results.[15]

CCD (1982; 1396) defined the *period 0 Malmquist input index*, $Q^0(x^1,x^0)$, for the production unit using the period 0 input distance function as follows:[16]

(5) $Q^0(x^1,x^0) \equiv D^0(y^0,x^1)/D^0(y^0,x^0) = D^0(y^0,x^1)$

where the last equality follows if production is efficient in period 0 since in this case, $D^0(y^0,x^0)$ equals one.[17] A value of the index greater than one implies that the input vector in period 1 is larger than the input vector in period 0, using the technology of period 0 as the reference technology.

The above input index depends only on the period 0 technology. Using the period 1 technology, CCD (1982; 1396) also defined the *period 1 Malmquist input index*, $Q^1(x^1,x^0)$, for the production unit as follows:

(6) $Q^1(x^1,x^0) \equiv D^1(y^1,x^1)/D^1(y^1,x^0) = 1/D^1(y^1,x^0)$

[15] In order to derive their results, CCD assumed that the first order partial derivatives of g^t and F^t existed at the observed period 0 and 1 data points. Thus in order to apply their results in the present context, we assume that the period 0 and 1 production possibility frontiers are differentiable at the observed data points.

[16] Note that (y^0,x^0) and (y^1,x^1) are the observed period 0 and 1 output and input vectors respectively for the production unit. We assume that all of these vectors are strictly positive.

[17] It should be mentioned at this point that throughout the paper, we assume that each observation is technically efficient. The reason for this somewhat restrictive assumption is that we want to apply index number techniques (rather than DEA techniques, which can readily deal with inefficiency) in order to obtain productivity growth decompositions. Index number methods cannot deal with technical inefficiency and this limitation must be kept in mind.

where the last equality follows if production is efficient in period 1 since in this case, $D^1(y^1,x^1)$ equals one. A value of the index greater than one implies that the input vector in period 1 is larger than the input vector in period 0, using the technology of period 1 as the reference technology.

Equations (5) and (6) define theoretical indexes which can be implemented empirically in alternative ways. For example, one way is to use linear programming techniques to estimate the distance functions.[18] An alternative is to derive index number formulae from the theoretical indexes. For example, Theorem 1 in CCD (1982; 1398) showed that the geometric mean of the two alternative input indexes (5) and (6) is numerically equal to the Törnqvist input index, Q_T defined by (8) below, provided that the production unit minimizes the cost of producing its observed output vector y^t in each period t (where it faces the vector of input prices $w^t >> 0_N$ in period t) and the input distance functions D^t have the translog functional form[19] where the quadratic term coefficients in the logarithms of the input vectors in $D^0(y,x)$ and $D^1(y,x)$ are identical; i.e., under these hypotheses we have:

$$(7)\ [Q^0(x^1,x^0)Q^1(x^1,x^0)]^{1/2} = Q_T(w^0,w^1,x^0,x^1)$$

where the logarithm of the *Törnqvist input index* Q_T is defined as follows:

$$(8)\ \ln Q_T(w^0,w^1,x^0,x^1) \equiv (1/2)\sum_{n=1}^{N} [s_n^{\ 0} + s_n^{\ 1}] \ln[x_n^{\ 1}/x_n^{\ 0}]$$

and the period t cost share for input n is defined as $s_n^{\ t} \equiv w_n^{\ t}x_n^{\ t}/w^t\cdot x^t$ for t = 0,1 and n = 1,...,N. CCD showed that this result holds without making any assumptions on returns to scale for the translog distance function, but as noted above, their result did require the assumption of competitive cost minimizing behaviour on the part of the producer.

CCD (1982; 1399-1401) derived analogous results for output indexes. CCD (1982; 1400) defined the *period 0 Malmquist output index*, $q^0(y^1,y^0)$, for the production unit using the period 0 output distance function as follows:[20]

$$(9)\ q^0(y^1,y^0) \equiv d^0(y^1,x^0)/d^0(y^0,x^0) = d^0(y^1,x^0)$$

where the last equality follows if production is efficient in period 0 since in this case, $d^0(y^0,x^0)$ equals one. A value of the index greater than one implies that the output vector in period 1 is larger than the output vector in period 0, using the technology of period 0 as the reference technology.

The above output index depends only on the period 0 technology. Using the period 1 technology, CCD (1982; 1400) also defined the *period 1 Malmquist output index*, $q^1(y^1,y^0)$, for the production unit as follows:

$$(10)\ q^1(y^1,y^0) \equiv d^1(y^1,x^1)/d^1(y^0,x^1) = 1/d^1(y^0,x^1)$$

where the last equality follows if production is efficient in period 1 since in this case, $d^1(y^1,x^1)$ equals one.

[18] See Färe, Grosskopf, Norris and Zhang (1994).

[19] For material on translog functional forms, see Christensen, Jorgenson and Lau (1973) and Diewert (1974).

[20] Note that (y^0,x^0) and (y^1,x^1) are the observed period 0 and 1 output and input vectors respectively for the production unit. We assume that all of these vectors are strictly positive.

Theorem 2 in CCD (1982; 1401) showed that the geometric mean of the two alternative input indexes defined by (9) and (10) is numerically equal to the Törnqvist output index, Q_T^* defined by (12) below, provided that the production unit maximizes the revenue it can raise conditional on using its observed input vector x^t in each period t (where it takes the period t vector of output prices $p^t >> 0_M$ as a vector of fixed parameters) and the output distance functions d^t have the translog functional form where the quadratic term coefficients in the logarithms of the output vectors in $d^0(y,x)$ and $d^1(y,x)$ are identical; i.e., under these hypotheses we have:

(11) $[q^0(y^1,y^0)q^1(y^1,y^0)]^{1/2} = Q_T^*(p^0,p^1,y^0,y^1)$

where the logarithm of the *Törnqvist output index* Q_T^* is defined as follows:

(12) $\ln Q_T^*(p^0,p^1,y^0,y^1) \equiv (1/2)\sum_{m=1}^{M} [S_n^0 + S_n^1] \ln[y_m^1/y_m^0]$

and the period t revenue share for output m is defined as $S_m^t \equiv p_m^t y_m^t / p^t \cdot y^t$ for $t = 0,1$ and $m = 1,...,M$. CCD showed that this result holds without making any assumptions on returns to scale for the translog distance function. However, their result required the assumption of competitive revenue maximizing behaviour on the part of the producer in each period, conditional on the observed vector of inputs, and this assumption may not be warranted in noncompetitive situations.

One approach to measuring productivity growth is to take ratios of Malmquist output indexes to Malmquist input indexes. Given two possible definitions for both input and output indexes, this leads to four possible productivity indexes. This approach was suggested by Hicks (1961)[21] and Moorsteen (1961), leading Diewert (1992) to label these as "Hicks-Moorsteen" indexes.

Consider taking the geometric means of the two alternative Malmquist input and output indexes, respectively, and then taking their ratios. When the distance functions have the translog functional form, as can be seen from equations (7) and (11), this corresponds to taking the ratio of a Törnqvist output index to a Törnqvist input index as a measure of productivity growth; this "standard" Törnqvist productivity index approach is used by the Bureau of Labor Statistics to construct their productivity estimates for the U.S. manufacturing sector. No assumptions need be made on the returns to scale of the underlying translog functional forms in order to derive this result; the invariance of the Malmquist input and output indexes to returns to scale assumptions was emphasized by CCD and also noted by Bjurek (1996). However, this approach to the measurement of productivity does require the assumption of competitive cost minimizing behaviour conditional on observed outputs and competitive revenue maximizing behaviour conditional on observed inputs, assumptions that may not be satisfied in many contexts.

[21] "This measure of input as a whole is not the same as the measure of output as a whole, as might perhaps be supposed at first sight. In the one case we should be asking whether A outputs could be produced from B inputs with B techniques; in the other whether B inputs would be sufficient to produce A outputs with A techniques; and vice versa for the other limb of the comparison. If all went well, the relation between the measure of output and the measure of input ought to give us a measure of the improvement in technique - or, as it might be better to say, a measure of the efficiency with which resources are combined on the one occasion compared with the other." J.R. Hicks (1961; 22).

In order to overcome the above limitations of the Hicks-Moorsteen-CCD approach to the measurement of productivity change, we will adapt another approach used by CCD (1982; 1401-1408) in order to obtain productivity growth indexes. Our suggested modification of their second approach allows us to derive a simple method of separating the contributions of technical change and returns to scale to productivity growth, without assuming price taking behavior for outputs.

3. Returns to Scale, Technical Change and Imperfect Competition

Using the technology of the producer at period 0, CCD (1982; 1402) defined a period 0 output based *productivity growth index* for the production unit going from period 0 to 1 as follows:

$$\begin{aligned}
(13)\ m^0(x^1,x^0,y^1,y^0) &\equiv d^0(y^1,x^1)/d^0(y^0,x^0) \\
&= d^0(y^1,x^1) && \text{if } d^0(y^0,x^0) = 1 \\
&= \min{}_{\delta>0}\ \{\delta : (y^1/\delta,x^1)\in S^0\} && \text{using definition (2)} \\
&= \delta^0.
\end{aligned}$$

We assume that (y^t,x^t) is on the frontier of the period t production possibilities set, S^t, for $t = 0,1$. Generally speaking, production possibilities sets grow over time so that S^0 will generally be a subset of S^1. In this case, given that (y^1,x^1) is on the frontier of S^1, it is likely that (y^1,x^1) will not belong to S^0. Hence we must in general deflate y^1 by a number larger than one so that the resulting deflated vector, y^1/δ, is just small enough so that $(y^1/\delta,x^1)$ will be on the frontier of S^0. This minimal deflation factor is δ^0, which will be equal to or greater than 1 if S^0 is a subset of S^1. Thus if $m^0(x^1,x^0,y^1,y^0) = \delta^0$ is greater than one, then we say that there has been *productivity growth* between the two periods. Using this productivity index, we are basically deflating the outputs of the period 1 production vector (y^1,x^1) so that the resulting deflated production vector $(y^1/\delta^0,x^1)$ is on the period 0 production surface.

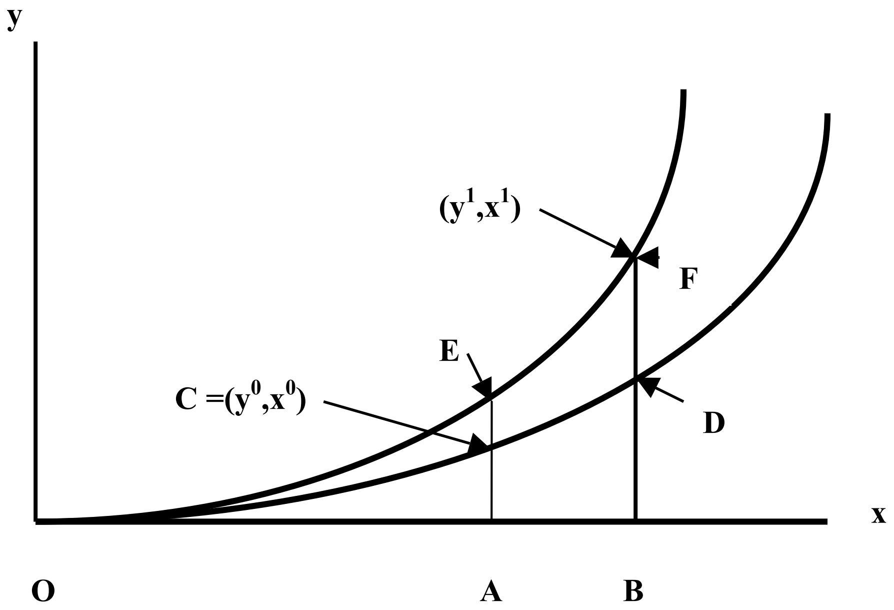
CCD (1982; 1402) defined the following companion period 1 output based *productivity growth index* for the production unit going from period 0 to 1 as follows:

$$\begin{aligned}
(14)\ m^1(x^1,x^0,y^1,y^0) &\equiv d^1(y^1,x^1)/d^1(y^0,x^0) \\
&= 1/d^1(y^0,x^0) && \text{if } d^1(y^1,x^1) = 1 \\
&= 1/\min{}_{\delta>0}\ \{\delta : (y^0/\delta,x^0)\in S^1\} && \text{using definition (2)} \\
&= \max{}_{\delta>0}\ \{\delta : (\delta y^0,x^0)\in S^1\} \\
&= \delta^1.
\end{aligned}$$

Assume that S^0 is a subset of S^1. In this case, given that (y^0,x^0) is on the frontier of S^0, it is unlikely that (y^0,x^0) will be on the frontier of S^1. Hence in order to obtain a production vector that is on the frontier of the set S^1, we must in general inflate y^0 by a number larger than one so that the resulting inflated vector, δy^0, is just large enough so that $(\delta y^0,x^0)$ will be on the frontier of S^1. This maximal inflation factor is δ^1, which will be equal to or greater than 1 if S^0 is a subset of S^1. Thus if $m^1(x^1,x^0,y^1,y^0) = \delta^1$ is greater than one, then CCD say that there has been *productivity growth* between the two periods. Using this productivity index, we are basically inflating the outputs of the period 0 production vector (y^0,x^0) so that the resulting inflated production vector $(\delta^1 y^1,x^0)$ is on the period 1 production surface.

In the case of one output and one input, it is easy to give a graphical interpretation of the above two CCD productivity indexes and we do this in Figure 1 below. The frontier of the period 0 production possibilities set is the line OCD (which is the period 0 production function) and the frontier of the period 1 production possibilities set is the line OEF (which is the period 1 production function). The observed output and input in period 0 is the point (y^0,x^0) (the point C) and the observed output and input in period 1 is the point (y^1,x^1) (the point F). Working through definitions (13) and (14) above, it can be verified that $m^0(x^1,x^0,y^1,y^0)$ is equal to the distance FB divided by the distance DB and $m^1(x^1,x^0,y^1,y^0)$ is equal to the distance EA divided by the distance CA.

Figure 1: The CCD Productivity Indexes in the One Output and One Input Case



It can be seen that the two CCD productivity indexes are not conventional productivity indexes, which are usually defined as an index of output growth divided by an index of input growth. In fact, m^0 and m^1 measure *shifts* in the production function going from period 0 to period 1. Thus in the time series context, m^0 and m^1 are actually measures of *technical progress*.[22] Hence in what follows, we will interpret these measures as technical progress measures.

As usual, there is no reason to prefer the technical progress index m^0 over the companion index m^1. Thus we can follow CCD (1982; 1404) and define an overall *Malmquist-CCD productivity growth index*, (or more accurately, a measure of *technical progress*), τ, as the geometric mean of the indexes m^0 and m^1:[23]

$$(15)\ \tau(x^1,x^0,y^1,y^0) \equiv [m^0(x^1,x^0,y^1,y^0)m^1(x^1,x^0,y^1,y^0)]^{1/2}.$$

We will follow the example of CCD and assume that the technologies of the producer in the two periods under consideration can be represented by the following two *translog output distance functions*, $d^t(y,x)$, for $t = 0,1$, where the logarithms of these functions are defined as follows for $t = 0,1$:

$$(16)\ \ln d^t(y,x) \equiv \alpha_0{}^t + \sum_{m=1}{}^M \alpha_m{}^t \ln y_m + \sum_{n=1}{}^N \beta_n{}^t \ln x_n + (1/2)\sum_{i=1}{}^N \sum_{j=1}{}^N \gamma_{ij} \ln x_i \ln x_j$$
$$+ (1/2)\sum_{k=1}{}^M \sum_{m=1}{}^M \delta_{km} \ln y_k \ln y_m + \sum_{m=1}{}^M \sum_{n=1}{}^N \phi_{mn} \ln y_m \ln x_n$$

where the parameters on the right hand side of (16) satisfy the following restrictions:

$$(17)\ \sum_{n=1}{}^N \beta_n{}^t < 0\ ;$$

$$(18)\ \sum_{j=1}{}^N \gamma_{ij} = 0 \text{ for } i = 1,...,N\ ;$$

$$(19)\ \gamma_{ij} = \gamma_{ji} \text{ for all } 1 \le i < j \le N\ ;$$

$$(20)\ \sum_{m=1}{}^M \alpha_m{}^t = 1\ ;$$

$$(21)\ \sum_{m=1}{}^M \delta_{km} = 0 \text{ for } k = 1,...,M\ ;$$

$$(22)\ \delta_{km} = \delta_{mk} \text{ for all } 1 \le k < m \le M\ ;$$

$$(23)\ \sum_{m=1}{}^M \phi_{mn} = 0 \text{ for } n = 1,...,N\ ;$$

$$(24)\ \sum_{n=1}{}^N \phi_{mn} = 0 \text{ for } m = 1,...,M.$$

Note that all of the quadratic parameters in the definitions of d^0 and d^1 are restricted to be the same in the two periods under consideration; only the constant term and linear terms are allowed to shift. Note also that the restrictions on the parameters (20)-(23) imply that each $d^t(y,x)$ is linearly homogeneous in the components of y; i.e., we have $d^t(\lambda y,x) = \lambda d^t(y,x)$ for all $\lambda > 0$ and $y >> 0_M$ and $x >> 0_N$, which is a property that output distance functions *must* satisfy. Technical progress is represented by changes in the constant term and the linear terms in definitions (16).

The two translog output distance functions defined by (16) and the following restrictions on the parameters given by (17)-(24) are *almost* completely flexible functional forms[24] for the case of a nonconstant returns to scale technology. However, imposing the restrictions

[22] This point did not emerge clearly in the exposition of CCD since they explained their indexes in a cross sectional context.

[23] In the one output, one input case, it can be seen that τ is the geometric average of the distances FB/DB and EA/CA.

[24] See Diewert (1974; 139) for materials on the flexibility of translog functional forms that involve two sets of variables.

(18) and (24) destroys this complete flexibility. We imposed these extra restrictions so that our measures of local returns to scale, defined below by (25), become constant parameters for each period.

Following CCD (1982; 1402), *local returns to scale* for the technology in period t, ε^t, can be defined using the derivatives of the period t output distance function d^t as follows for t = 0,1:[25]

$$\begin{aligned}
(25)\ \varepsilon^t &\equiv -[\partial d^t(y^t,\lambda x^t)/\partial\lambda_{\,|\,\lambda=1}] \\
&= -\ [x^t\cdot\nabla_x d^t(y^t,x^t)] \\
&= -\ [\partial \ln d^t(y^t,\lambda x^t)/\partial\lambda_{\,|\,\lambda=1}] && \text{using } d^t(y^t,x^t) = 1 \\
&= -\ [\textstyle\sum_{n=1}^{N} \partial \ln d^t(y^t,x^t)/\partial \ln x_n] \\
&= -\ \textstyle\sum_{n=1}^{N} \beta_n^{\,t} && \text{differentiating (16) and using (18), (19) and (24)} \\
&> 0 && \text{using (17)}
\end{aligned}$$

where $\nabla_x d^t(y^t,x^t)$ is the vector of derivatives of $d^t(y^t,x^t)$ with respect to the components of x. Thus (25) tells us that the degree of returns to scale in period t, ε^t, is a positive constant, which is equal to minus the sum of the $\beta_n^{\,t}$ parameters which match up with the $\ln x_n$ variables in the definition of the period t output distance function d^t defined by (16).[26]

Assuming that the producer's technology in each period can be represented by the translog output distance functions defined by (16)-(24), we can now work out expressions for the technical progress indexes, m^0, m^1 and τ defined by (13)-(15). To do this, we assume (for the remainder of the paper) that production is efficient in each period so that:

$$(26)\ d^0(y^0,x^0) = d^1(y^1,x^1) = 1.$$

Definition (14) and assumptions (26) imply that

$$(27)\ m^1(x^1,x^0,y^1,y^0) \equiv d^1(y^1,x^1)/d^1(y^0,x^0) = d^0(y^0,x^0)/d^1(y^0,x^0).$$

Taking logarithms of both sides of (27) and using (16), we find that

$$\begin{aligned}
(28)\ \ln m^1(x^1,x^0,y^1,y^0) &= \ln d^0(y^0,x^0) - \ln d^1(y^0,x^0) \\
&= \alpha_0^{\,0} - \alpha_0^{\,1} + \textstyle\sum_{m=1}^{M} (\alpha_m^{\,0} - \alpha_m^{\,1})\ln y_m^{\,0} + \sum_{n=1}^{N} (\beta_n^{\,0} - \beta_n^{\,1})\ln x_n^{\,0}.
\end{aligned}$$

Definition (13) and assumptions (26) imply that

$$(29)\ m^0(x^1,x^0,y^1,y^0) \equiv d^0(y^1,x^1)/d^0(y^0,x^0) = d^0(y^1,x^1)/d^1(y^1,x^1).$$

[25] In order to justify definition (25), we require that $d^t(y^t,x^t) = 1$ so that the observed period t production vector is efficient. Note that $\varepsilon^t > 1$, $\varepsilon^t = 1$ and $\varepsilon^t < 1$ means that there are increasing, constant or decreasing returns to scale in period t respectively.

[26] The translog distance functions defined by (16)-(24) for each period are completely flexible functional forms in the class of constant returns to scale technologies. In this case, it turns out that $d^t(y,x)$ must be homogeneous of degree −1 in the components of x; i.e., d^t must satisfy $d^t(y,\lambda x) = \lambda^{-1}d^t(y,x)$ for all $\lambda > 0$, $y \gg 0_M$ and $x \gg 0_N$. This extra homogeneity condition can be imposed upon the d^t defined by (16)-(24) if we replace the restriction (17) by the restriction $\sum_{n=1}^{N} \beta_n^{\,t} = -1$. Thus our more general restriction (17) adds an extra free parameter to our specification and allows general nonconstant returns to scale in a very parsimonious way (and of course, constant returns to scale is allowed as a special case of our specification).

Taking logarithms of both sides of (29) and using (16), we find that

$$(30)\ \ln m^0(x^1,x^0,y^1,y^0) = \ln d^0(y^1,x^1) - \ln d^1(y^1,x^1)$$
$$= \alpha_0{}^0 - \alpha_0{}^1 + \textstyle\sum_{m=1}{}^M (\alpha_m{}^0 - \alpha_m{}^1)\ln y_m{}^1 + \sum_{n=1}{}^N (\beta_n{}^0 - \beta_n{}^1)\ln x_n{}^1.$$

Later in the paper, we will assume that the linear coefficients in the two translog functions defined by (16) are all equal so that we have:

$$(31)\ \alpha_m{}^0 = \alpha_m{}^1\ ;\ m = 1,...,M\ ;\ \beta_n{}^0 = \beta_n{}^1\ ;\ n = 1,...,N.$$

If assumptions (31) are satisfied, then definition (15) and equations (28) and (30) imply that all of our measures of technical progress are equal to the same constant; i.e., we have

$$(32)\ \tau(x^1,x^0,y^1,y^0) = m^0(x^1,x^0,y^1,y^0) = m^1(x^1,x^0,y^1,y^0) = \exp(\alpha_0{}^0 - \alpha_0{}^1) \equiv \tau^*.$$

Thus under assumptions (31), there will be *positive technical progress* going from the period 0 technology to the period 1 technology in our translog model provided that τ^* is greater than one and this condition will hold if and only if $\alpha_0{}^0 - \alpha_0{}^1$ is greater than 0.

As was mentioned above, a difficulty with the CCD methodology is that it assumed competitive revenue maximizing behavior on the part of the producer, conditional on the observed input vector x^t in each period. However, if there are increasing returns to scale in each period so that ε^0 and ε^1 are greater than one, then it is well known that competitive profit maximizing behavior breaks down. Thus for each period t, we assume that the firm or production unit faces the *inverse demand functions* $P_m{}^t(y_m)$ which give the market clearing prices for output m as a function of the amount of output y_m that the firm produces, for m = 1,…,M. Assuming that the firm faces the strictly positive input price vector $w^t \equiv [w_1{}^t,\ldots,w_N{}^t]$ in period t, the *firm's period t monopolistic profit maximization problem* is the following constrained maximization problem involving the vector of period t outputs $y \equiv [y_1,\ldots,y_M]$ and the input vector x:

$$(33)\ \max{}_{y,x}\ \{\textstyle\sum_{m=1}{}^M P_m{}^t(y_m)y_m - w^t\cdot x : (y,x)\in S^t\}\ ; \qquad t = 0,1.$$

We assume that that for t = 0,1, the strictly positive period t observed output and input vectors, y^t and x^t, solve the period t monopolistic profit maximization problem and that the *observed period t prices for the outputs* are:[27]

$$(34)\ p_m{}^t \equiv P_m{}^t(y_m{}^t)\ ; \qquad m = 1,\ldots,M\ ;\ t = 0,1.$$

Assuming that the demand derivatives $dP_m{}^t(y_m{}^t)/dy_m$ are nonpositive, the nonnegative *ad valorem monopolistic markup* $\mu_m{}^t$ for the mth output in period t can be defined as follows:

$$(35)\ \mu_m{}^t \equiv -\ [dP_m{}^t(y_m{}^t)/dy_m][y_m{}^t/p_m{}^t] \geq 0\ ; \qquad m = 1,\ldots,M\ ;\ t = 0,1.$$

CCD assumed that y^t and x^t were solutions to certain (competitive) revenue maximization and cost minimization problems.[28] We need to develop noncompetitive counterparts to these assumptions made by CCD. In order to accomplish this task, we note that our

[27] We assume that the functions $P_m{}^t(y_m)$ are differentiable around $y_m{}^t$. If the production unit has constant or decreasing returns to scale and behaves competitively, then this case can be modeled by setting $P_m{}^t(y_m)$ equal to the constant output price $p_m{}^t$ for m = 1,...,M and t = 0,1.

[28] See equations (25) and (37) in CCD.

assumption that (y^t,x^t) solves the period t monopolistic profit maximization problem defined by (33) for t = 0,1 means that the following equalities are satisfied:

$$
\begin{aligned}
(36)\ & \max_{y,x} \{\textstyle\sum_{m=1}^{M} P_m^t(y_m)y_m - w^t\cdot x : (y,x)\in S^t\} \qquad t = 0,1 \\
& = \max_{y} \{\textstyle\sum_{m=1}^{M} P_m^t(y_m)y_m - w^t\cdot x^t : (y,x^t)\in S^t\} \\
& = \max_{y} \{\textstyle\sum_{m=1}^{M} P_m^t(y_m)y_m : (y,x^t)\in S^t\} - w^t\cdot x^t \\
& = \max_{x} \{\textstyle\sum_{m=1}^{M} P_m^t(y_m^t)y_m^t - w^t\cdot x : (y^t,x)\in S^t\} \\
& = \textstyle\sum_{m=1}^{M} P_m^t(y_m^t)y_m^t - \min_{x} \{w^t\cdot x : (y^t,x)\in S^t\}.
\end{aligned}
$$

Thus for t = 0,1, y^t is a solution to the following *conditional on* x^t *monopolistic revenue maximization problem*:

$$
\begin{aligned}
(37)\ \textstyle\sum_{m=1}^{M} P_m^t(y_m^t)y_m^t & = \max_{y} \{\textstyle\sum_{m=1}^{M} P_m^t(y_m)y_m : (y,x^t)\in S^t\} \qquad t = 0,1 \\
& = \max_{y} \{\textstyle\sum_{m=1}^{M} P_m^t(y_m)y_m : x_1^t = g^t(y,\tilde{x}^t)\}
\end{aligned}
$$

where we have used the period t input requirements function g^t to represent the technology constraints in the second maximization problem in (36) instead of using the production possibilities set S^t. This second maximization problem is the counterpart to the maximization problem (25) in CCD (1982; 1400). Assuming that g^t is differentiable when evaluated at the period t data, the following first order necessary conditions for maximizing (37) must be satisfied:

$$(38)\ p_m^t\,(1 - \mu_m^t) = \lambda^t\, \partial g^t(y^t,\tilde{x}^t)/\partial y_m\ ; \qquad m = 1,\ldots,M\ ;\ t = 0,1$$

where the μ_m^t are defined by (35). Now multiply equation m in (38) for period t by y_m^t, sum the resulting equations over m, solve for the period t Lagrange multiplier λ^t and substitute the resulting expression for λ^t back into equations (38). The resulting equations are:

$$
\begin{aligned}
(39)\ p_m^t\,(1 - \mu_m^t)/[\textstyle\sum_{k=1}^{M} p_k^t\,(1 - \mu_k^t)y_k^t] & = [\partial g^t(y,\tilde{x}^t)/\partial y_m]/y^t\cdot\nabla_y g^t(y^t,\tilde{x}^t) \\
& = \partial d^t(y^t,x^t)/dy_m\ ; \qquad m = 1,\ldots,M\ ;\ t = 0,1
\end{aligned}
$$

where the second set of equalities in (39) follows from a general result established by CCD (1982; 1399).

If the individual product markups happen μ_m^t happen to be equal to a common markup μ^t in each period,[29] or if there is only one output, then it can be seen that conditions (39) collapse down to the following simpler conditions:

$$(40)\ p^t/p^t\cdot y^t = \nabla_y d^t(y^t,x^t)\ ; \qquad t = 0,1.$$

Recalling our assumption that (y^t,x^t) solves the period t monopolistic profit maximization problem defined by (33) for t = 0,1, the fourth equality in (36) implies that the observed period t input vector x^t is a solution to the following *period t conditional on* y^t *cost minimization problem*:

$$(41)\ w^t\cdot x^t = \min_{x} \{w^t\cdot x : (y^t,x)\in S^t\}\ ; \qquad t = 0,1$$

[29] The case of competitive price taking behavior is a special case where $\mu^t = 0$ for t = 0,1.

$$= \min_{x} \{\textstyle\sum_{n=1}^{N} w_n^t x_n : x_1 = g^t(y^t,x_2,\ldots,x_N)\}.$$

Assuming that $g^t(y^t,x_2,\ldots,x_N)$ is once differentiable with respect to $x_2,\ldots,x_N$, the first order necessary conditions for the period t cost minimization problems represented by the second equation in (41) will hold and we can repeat the algebra developed by CCD (1982; 1403-1404) and show that the derivatives of the period t output distance function with respect to the components of x exist when evaluated at $x = x^t$ and have the following form:

$$(42)\ \nabla_x d^t(y^t,x^t) = -\ \varepsilon^t\, w^t/w^t\cdot x^t\ ; \qquad\qquad t = 0,1$$

where ε^t is the period t degree of local returns to scale defined by (25) above. If the production unit's distance function is defined by (16), then (25) shows that $-\varepsilon^t$ is equal to the sum of the β_n^t parameters in definitions (16) for t = 0,1.

Now we are ready to establish a monopolistic competition version of CCD's (1982; 1407-1408) Theorem 4. We assume that the firm's output distance function d^t in each period t has the translog functional form defined by (16)-(24). We also assume that production is efficient in each period so that conditions (26) hold.

Recall that the Malmquist-CCD technical progress index, τ, was defined by (15). Taking logarithms of both sides of (15) and using definitions (13) and (14) and assumptions (16), we have for t = 0,1:

$$(43)\ \ln \tau(x^1,x^0,y^1,y^0) = (1/2)[\ln d^0(y^1,x^1) - \ln d^0(y^0,x^0)] + (1/2)[\ln d^1(y^1,x^1) - \ln d^1(y^0,x^0)]$$

$$= (1/4)[\nabla_{\ln y} \ln d^0(y^1,x^1) + \nabla_{\ln y} \ln d^0(y^0,x^0)]\cdot[\ln y^1 - \ln y^0]$$

$$+ (1/4)[\nabla_{\ln x} \ln d^0(y^1,x^1) + \nabla_{\ln x} \ln d^0(y^0,x^0)]\cdot[\ln x^1 - \ln x^0]$$

$$+ (1/4)[\nabla_{\ln y} \ln d^1(y^1,x^1) + \nabla_{\ln y} \ln d^1(y^0,x^0)]\cdot[\ln y^1 - \ln y^0]$$

$$+ (1/4)[\nabla_{\ln x} \ln d^1(y^1,x^1) + \nabla_{\ln x} \ln d^1(y^0,x^0)]\cdot[\ln x^1 - \ln x^0]$$

using (16) and applying Diewert's (1976; 118) quadratic identity twice

$$= (1/2)[\nabla_{\ln y} \ln d^0(y^0,x^0) + \nabla_{\ln y} \ln d^1(y^1,x^1)]\cdot[\ln y^1 - \ln y^0]$$

$$+ (1/2)[\nabla_{\ln x} \ln d^0(y^0,x^0) - \nabla_{\ln x} \ln d^1(y^1,x^1)]\cdot[\ln x^1 - \ln x^0]$$

using (16) and applying CCD's (1982; 1404) generalized translog identity[30]

$$= \ln Q_T^{*}(p_1^0(1-\mu_1^0),\ldots,p_M^0(1-\mu_M^0);p_1^1(1-\mu_1^1),\ldots,p_M^1(1-\mu_M^1);y^0,y^1)$$

$$- (1/2)[(\varepsilon^0 w^0/w^0\cdot x^0) + (\varepsilon^1 w^1/w^1\cdot x^1)]\cdot[\ln x^1 - \ln x^0] \qquad \text{using (39) and (42)}$$

[30] A referee asked whether similar results hold for other functional forms. Analogous exact index number results do hold for other functional forms but typically, the results are messier than the comparable results for the translog functional form; see Diewert (2002) (2009) for a listing of superlative index number formulae. All of these exact and superlative index number results rely on the underlying functional form being quadratic or a simple transformation of a quadratic functional form since the main tool used to derive the exact index number formulae is Diewert's (1976; 118) Quadratic Identity and the Translog Identity in CCD (1982; 1412), which is a generalization of the Quadratic Identity. The translog functional form works well in this context because it is easy to impose restrictions on the parameters that ensure that the translog functional form has appropriate homogeneity properties.”

where the Törnqvist output index $Q_T^*(p^0,p^1,y^0,y^1)$ was defined by (12) above but in the above application, the observed output prices for period t, $p^t \equiv [p_1^t,...,p_M^t]$ are replaced by the *output prices adjusted for monopolistic markups*, $[p_1^t(1-\mu_1^t),...,p_M^t(1-\mu_M^t)]$ where the ad valorem markups μ_m^t are defined by (35) above for t = 0,1 and m = 1,...,M.

The final equation in (43) can be simplified if we define the period t vector of *marginal costs*, π^t, as follows:

(44) $\pi^t \equiv [\pi_1^t,...,\pi_M^t] \equiv [p_1^t(1-\mu_1^t),...,p_M^t(1-\mu_M^t)]$; t = 0,1.

To see why the π^t vectors can be interpreted as vectors of marginal costs, define the *firm's period t cost function*, c^t, as follows:

(45) $c^t(y,w) \equiv \min_x \{w\cdot x : (y,x)\in S^t\}$; t = 0,1.

From equations (36), it can be seen that our assumptions imply that the observed period t output vector, y^t, is a solution to the following period t monopolistic profit maximization problem:

(46) $\max_y \{\sum_{m=1}^M P_m^t(y_m)y_m - c^t(y,w^t)\}$; t = 0,1.

Assuming that $c^t(y,w^t)$ is differentiable with respect to the components of y at $y = y^t$, the first order necessary conditions for (46) imply the following conditions:

(47) $p_m^t(1 - \mu_m^t) = \partial c^t(y^t,w^t)/\partial y_m$; m = 1,...,M ; t = 0,1.

Using definitions (44), conditions (47) can be written more succinctly as $\pi^t = \nabla_y c^t(y^t,w^t)$ for t = 0,1.

Using definitions (44), (43) can be rewritten as follows:

(48) $\ln\tau(x^1,x^0,y^1,y^0)$

$$= \ln Q_T^*(\pi^0,\pi^1,y^0,y^1) - (1/2)[(\varepsilon^0 w^0/w^0\cdot x^0) + (\varepsilon^1 w^1/w^1\cdot x^1)]\cdot[\ln x^1 - \ln x^0].$$

The above equation is the main result in this paper. Thus we have the following result:[31]

Proposition 1: Suppose that the technology of a production unit can be represented by the translog output distance functions defined by (16)-(24) for periods 0 and 1.[32] Suppose further that $(y^t,x^t) >> 0_{M+N}$ solves the monopolistic profit maximization problem (33) for t = 0,1 where the inverse demand functions $P_m^t(y_m)$ are differentiable at y_m^t with $dP_m^t(y_m^t)/dy_m \le 0$ for m = 1,...,M and t = 0,1. Then the logarithm of the Malmquist-CCD productivity (or more accurately, technical progress) index, $\tau(x^1,x^0,y^1,y^0)$ defined by (15), is equal to the right hand side of (48) where the vector of marginal cost prices π^t is defined by (44) and (35) and the degree of local returns to scale at the period t data, ε^t, is defined by (25) for t = 0,1.

[31] This result is similar to a result obtained by Diewert and Fox (2008; 178) except that they used translog cost functions instead of translog distance functions in order to obtain their main result.

[32] In order to apply various results in CCD, we also require that the period t output distance function, $d^t(y,x)$, be locally dual to a differentiable input requirements function, $x_1 = g^t(y,\tilde{x})$, around the point y^t,x^t with $y^t\cdot\nabla_y g^t(y^t,\tilde{x}^t) > 0$ for t = 0,1.

Corollary 1: Suppose that the α_m^t and β_n^t parameters in the two translog distance functions do not depend on time so that assumptions (31) hold. In this case, the degree of local returns to scale is constant across time so that we have $-\sum_{n=1}^N \beta_n = \varepsilon^0 = \varepsilon^1 \equiv \varepsilon$ and technical progress is also constant so that, recalling (32), we have $\tau(x^1,x^0,y^1,y^0) = \exp(\alpha_0^0 - \alpha_0^1) \equiv \tau^*$. Under these conditions, equation (48) simplifies to:

(49) $\ln\tau^* = \ln Q_T^*(\pi^0,\pi^1,y^0,y^1) - \varepsilon \ln Q_T(w^0,w^1,x^0,x^1)$

where $\ln Q_T(w^0,w^1,x^0,x^1)$ is the logarithm of the Törnqvist input index defined earlier by (8).

Corollary 2: Suppose that in addition to assumptions (31), all of the ad valorem markups are equal in each period, or there is only one output. In this case, the period t marginal cost price vectors π^t in the Törnqvist output index can be replaced by the observed period t output prices p^t and (49) simplifies to:

(50) $\ln\tau^* = \ln Q_T^*(p^0,p^1,y^0,y^1) - \varepsilon \ln Q_T(w^0,w^1,x^0,x^1)$.

Corollary 3: Suppose in addition to the restrictions (31) we have constant returns to scale in production (so that $-\sum_{n=1}^N \beta_n = 1 = \varepsilon$) and price taking behavior on the part of the producer in each period (so that each $\mu_m^t = 0$). Then (49) simplifies to:

(51) $\tau^* = Q_T^*(p^0,p^1,y^0,y^1)/Q_T(w^0,w^1,x^0,x^1)$;

i.e., technical progress, τ^*, is equal to the Törnqvist output index divided by the Törnqvist input index, which is the conventional Total Factor Productivity growth index used by Jorgenson and Griliches (1967) in their pioneering study.[33]

Corollary 3 provides an exact (and superlative) index number justification for the productivity index introduced by Jorgenson and Griliches (1967).

Note that (49) can be rewritten as follows:

(52) $\ln Q_T^*(\pi^0,\pi^1,y^0,y^1) = \ln\tau^* + \varepsilon \ln Q_T(w^0,w^1,x^0,x^1)$.

Equation (52) can be used as an equation that explains aggregate output growth; i.e., the logarithm of an output index, $\ln Q_T^*(\pi^0,\pi^1,y^0,y^1)$, is “explained” by technical change, $\ln\tau^*$, plus the logarithm of input growth, $\ln Q_T(w^0,w^1,x^0,x^1)$, except that this input growth term is multiplied by the degree of returns to scale, ε. If there are increasing returns to scale so that ε is greater than one and if there is input growth so that Q_T is greater than one and hence $\ln Q_T$ is greater than zero, then the input growth term, $\ln Q_T$, is *magnified* by the increasing returns to scale term, leading to a greater rate of output growth than can be explained by simply adding up input growth and technical progress. However, in order to implement this growth decomposition, we generally need to have some knowledge of the marginal cost prices in the two periods, π^0 and π^1. Of course, if all of the ad valorem markups are the same in each period or there is only one output, then $\ln Q_T^*(\pi^0,\pi^1,y^0,y^1)$ can be replaced by $\ln Q_T^*(p^0,p^1,y^0,y^1)$ and then the resulting equation (52) extended to many periods could be used as the starting point for an econometric specification that

[33] Jorgenson and Griliches derived their productivity index using a continuous time Divisia type approach rather than using a discrete time approach as is done here.

would estimate the unknown parameters τ^* and ε.[34] The error terms that result from econometrically estimating this model could be interpreted as unexplained productivity growth effects.[35] That is, in each period, the error term could be considered as a productivity shock unexplained by returns to scale and smooth rates of technical change. In many macroeconomic models, it is productivity shocks such as these which are of interest, rather than secular productivity growth driven by returns to scale and smooth rates of technical progress.

4. Conclusion

As indicated in the introduction, there is a considerable amount of theoretical interest in determining whether the Caves, Christensen and Diewert (1982) economic approach to obtaining productivity indexes is consistent with a distance function approach to the measurement of productivity change. The distance function approach to the measurement of productivity change can be implemented without making any assumptions about pricing behavior, which is an advantage of this approach. On the other hand, CCD showed how distance function measures of productivity growth could be estimated empirically using fairly simple index numbers (augmented by exogenous estimates of returns to scale) provided one made some assumptions about pricing behavior. Our conclusion is that the CCD approach is not fully satisfactory because their assumptions about producer behavior are not plausible in the case where there are increasing returns to scale. In the present paper, we modify their assumptions about producer behavior by assuming that the observed price and quantity data are consistent with a monopolistic profit maximizing model and we rework the analysis of CCD in order to obtain a variant of their results. This variant is equation (48) in the previous section or the simplified version of (48) that assumes that the degree of returns to scale is the same in each period, which is (52). Unfortunately, these new equations are more complicated than the corresponding equation in Caves, Christensen and Diewert (1982; 1404): in the present model (in the general case), the observed output prices p^t which appear in CCD must be replaced by difficult to observe marginal cost prices π^t. This will limit the usefulness of the present framework but it does have the benefit of being logically consistent when the

[34] Finding an appropriate econometric specification is not a trivial problem due to endogeneity problems. The input price vectors, w^0 and w^1, can be regarded as exogenous but the output and input vectors, y^t and x^t, and the selling price vectors p^t for $t = 0,1$, are all endogenous variables. Econometric issues in similar regression models are discussed by Bartelsman (1995), Burnside (1996), Basu and Fernald (1997) and Diewert and Fox (2008).

[35] This sentence requires a bit of elaboration. If we somehow know all of the price and quantity vectors that appear in equation (52), when we extend the analysis from 2 periods to T+1 periods, we will end up with T technical progress parameters of the form $\ln\tau^*$ and one returns to scale parameter ε. But we will have only T degrees of freedom to estimate these T+1 parameters. Thus it is natural to introduce an econometric model that assumes that these technical progress parameters behave in a "smooth" manner; i.e., a constant rate or linear or quadratic trends or linear spline trends in the τ^t. Then the residuals in the resulting regression model can be interpreted as deviations from the smoothed period to period rates of technical progress or these residuals could be interpreted as technical progress "shocks".

underlying technologies exhibit increasing returns to scale.[36] In general, marginal costs can be estimated through econometric,[37] engineering or accounting studies. In the special cases where there is only one output or where ad valorem markups can be assumed to be the same in each period across outputs, our new framework essentially reduces to the CCD model.[38]

An alternative to the economic approach to productivity measurement (which is the approach taken in this paper) is the axiomatic approach. The axiomatic approach works as follows: choose a functional form for a quantity index, say $Q^*(p^0,p^1,y^0,y^1)$ for the output index and $Q(w^0,w^1,x^0,x^1)$ for the input index. These choices of functional form are determined on the basis of the test or axiomatic approach to index number theory.[39] Then the *axiomatic productivity index* $A(p^0,p^1,y^0,y^1,w^0,w^1,x^0,x^1)$ is simply defined as the output index divided by the input index:

$$(53)\ A(p^0,p^1,y^0,y^1,w^0,w^1,x^0,x^1) \equiv Q^*(p^0,p^1,y^0,y^1)/Q(w^0,w^1,x^0,x^1).$$

Note that observed market prices are used as the price weights in the above quantity indexes. Now it is certainly true that the above axiomatic approach to measuring productivity growth can be consistent with the economic approach since Corollary 3 in the previous section shows that (53) is justified from the viewpoint of the economic approach (under certain conditions) if we choose Q^* and Q to be Törnqvist indexes. However, if there is noncompetitive behavior in the pricing of outputs on the part of producers, the analysis in the previous section shows that the axiomatic approach is not necessarily consistent with the economic approach. In particular, in noncompetitive contexts, from the viewpoint of the economic approach, it is not generally appropriate to use observed output prices in the output quantity index; instead marginal cost weights should be used. Thus if there is a discrepancy between the axiomatic and economic approach to the measurement of productivity growth, a certain amount of caution should be used in interpreting the axiomatic results.

Appendix: Distance Functions and Regularity Conditions on the Technology

Recall definitions (1) and (2) in the main text which defined the input and output distance functions, $D^t(y,x)$ and $d^t(y,x)$, which corresponded to the technology set S^t. In this Appendix, we will place restrictions on the sets S^t which are sufficient to ensure that the maximum in definition (1) and the minimum in definition (2) exist and are finite, provided that the output and input vectors, y and x, are strictly positive.

[36] Our new framework will also be useful in situations where there are constant returns to scale in production but innovative new technologies are developed and producers behave in a monopolistic manner. Our framework will also be useful in regulatory contexts where selling prices are set by the regulator but these selling prices are not equal to marginal costs.

[37] See Diewert and Lawrence (2005) for an example of econometric model where markups are estimated in a flexible functional form model.

[38] CCD did not work out the restrictions on the translog distance functions that make returns to scale constant over time periods.

[39] Fisher (1922) was a pioneer in this area of research. For more recent material on the axiomatic approach, see Diewert (1992) and Balk (1995).

In order to simplify the notation, we will drop the superscript t in what follows. We assume that the production possibilities set S is given and for $y >> 0_M$ and $x >> 0_N$, the *input distance function* D and the *output distance function* d are defined as follows:

(A1) $D(y,x) \equiv \max_{\delta>0} \{\delta : (y,x/\delta)\in S\}$.

(A2) $d(y,x) \equiv \min_{\delta>0} \{\delta : (y/\delta,x)\in S\}$.

Consider the following *four properties for S*:

P1. S is a nonempty closed subset of the nonnegative orthant in Euclidean M+N dimensional space.

P2. For every $y \geq 0_M$, there exists an $x \geq 0_N$ such that $(y,x)\in S$.

The interpretation of P2 is that every finite output vector y is producible by a finite input vector x.

P3. $(y,x^1)\in S$, $x^2 \geq x^1$ implies $(y,x^2)\in S$.

Thus if S satisfies P3, then there is free disposability of inputs.

P4. $y > 0_M$ implies that $(y,0_N)\notin S$.

The interpretation of P4 is that zero amounts of all inputs cannot produce a positive output.

We can now prove the following Proposition:

Proposition 2: Let $y > 0_M$ and $x >> 0_N$. Then D(y,x) is well defined as the maximum in (A1) with $D(y,x) > 0$ provided that S satisfies properties P1-P4.

Proof: Let $y > 0_M$ and $x >> 0_N$. Then by P2, there exists $x^* \geq 0_N$ such that $(y,x^*)\in S$. Since $x >> 0_N$, there exists a $\delta^* > 0$ that is small enough such that $x/\delta^* \geq x$. Thus by P3, $(y,x/\delta^*)\in S$. We cannot increase δ^* to plus infinity and conclude that $(y,0_N)\in S$ because this would contradict P4. Using the fact that S is a closed set , it can be seen that the maximization problem defined by (A1) has a finite positive maximum, δ^{**}. Q.E.D.

In order to show that the output distance function d(y,x) defined by (A2) is well defined as a positive minimum, we will require an additional three properties that S must satisfy:

P5. $x \geq 0_N$, $(y,x)\in S$ implies $0_M \leq y \leq b(x)1_M$ where 1_M is a vector of ones of dimension M and $b(x) \geq 0$ is a finite nonnegative bound.

The interpretation of P5 is: bounded inputs imply bounded outputs.

P6. $x >> 0_N$ implies that there exists $y >> 0_M$ such that $(y,x)\in S$.

Thus the technology is such that every strictly positive input vector can produce a strictly positive vector of outputs.

P7. $(y^1,x)\in S$, $0_M \leq y^2 \leq y^1$ implies $(y^2,x)\in S$.

Thus if the input vector x can produce the output vector y^1 and y^2 is equal to or less than y^1, then x can also produce the smaller vector of outputs, y^1 (free disposability of outputs).

Proposition 3: Let $y >> 0_M$ and $x >> 0_N$. Then d(y,x) is well defined as the minimum in (A2) with d(y,x) > 0 provided that S satisfies properties P1 and P5-P7.

Proof: Let $y >> 0_M$ and $x >> 0_N$. Since $x >> 0_N$, by P6, there exists a $y^* >> 0_M$ such that $(y^*,x)\in S$. Since y^* and y are strictly positive, there exists $\delta^* > 0$ large enough so that $y/\delta^* \leq y^*$. Using P7, we see that $(y/\delta^*,x)\in S$ and thus we have a feasible solution for the minimization problem in (A2). From definition (A2), we want to make $\delta \geq 0$ as small as possible such that $(y/\delta,x)\in S$. However, we cannot make $\delta > 0$ but arbitrarily close to 0 and have $(y/\delta,x)$ belong to S because this would contradict property P5. Using property P1, we see that a finite positive minimum for the minimization problem in (A2) exists. Q.E.D.

References

Alam, I.M.S. (2001), “A Nonparametric Approach for Assessing Productivity Dynamics of Large U.S. Banks”, *Journal of Money, Credit and Banking* 33, 121-139.

Atkinson, S.E., C. Cornwell and O. Honerkamp (2003), “Measuring and Decomposing Productivity Change: Stochastic Distance Function Estimation versus Data Envelopment Analysis”, *Journal of Business and Economic Statistics* 21, 284-294.

Balk, B.M. (1995), “Axiomatic Price Index Theory: A Survey”, *International Statistical Review* 63, 69-93.

Balk, B. M. (1998), *Industrial Price, Quantity, and Productivity Indices: The Micro-Economic Theory and an Application*, Boston/Dordrecht/London: Kluwer Academic Publishers.

Balk, B.M. (2001), “Scale Efficiency and Productivity Change”, *Journal of Productivity Analysis* 15, 159-183.

Bartelsman, E.J. (1995), “Of Empty Boxes: Returns to Scale Revisited”, *Economics Letters* 49, 59-67.

Basu, S. and J.G. Fernald (1997), “Returns to Scale in U.S. Production: Estimates and Implications”, *Journal of Political Economy* 105, 249-283.

Bennett, R.L. and R.E.A. Farmer (2003), “Indeterminacy with Non-Separable Utility”, *Journal of Economic Theory* 93, 118-143.

Benhabib, J. and Y. Wen (2004), “Indeterminancy, Aggregate Demand, and the Real Business Cycle”, *Journal of Monetary Economics*, 51, 503-530.

Bjurek, H. (1996), “The Malmquist Total Factor Productivity Index”, *Scandinavian Journal of Economics* 98, 303-313.

Bureau of Labor Statistics (2002), “Summary of Methods for the Manufacturing Sector and Manufacturing Industries”, United States Department of Labor, Washington D.C. Web address: www.bls.gov/news.release/prod5.tn.htm

Briec, W. and K. Kristiaan (2004), "A Luenberger-Hicks-Moorsteen Productivity Indicator: Its relation to the Hicks-Moorsteen Productivity Index and the Luenberger Productivity Indicator", *Economic Theory* 23, 925-939.

Burnside, C. (1996), "Production Function Regressions, Returns to Scale, and Externalities", *Journal of Monetary Economics* 37, 177-201.

Caves, D.W., L.R. Christensen and W.E. Diewert (1982), "The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity", *Econometrica* 50, 1393-1414.

Christensen, L.R., D.W. Jorgenson and L.J. Lau (1973), "Transcendental Logarithmic Production Frontiers", *Review of Economics and Statistics* 55, 28-45.

Ciccone, A. (2002), "Input Chains and Industrialization", *Review of Economic Studies* 69, 565-587.

Coelli, T., D.S.P. Rao and G. Battese (1998), *An Introduction to Efficiency and Productivity Analysis*, Boston/Dordrecht/London: Kluwer Academic Publishers.

Coelli, T., A. Estache, S. Perelman and L. Trujillo (2003), *A Primer on Efficiency Measurement for Utilities and Transport Regulators*, WBI Development Studies, Washington, D.C.: World Bank.

Cooper, W.W., L.M. Seiford and J. Zhu (eds.) (2004), *Handbook on Data Envelopment Analysis*, Boston/Dordrecht/London: Kluwer Academic Publishers.

De Borger, B. and K. Kristiaan (2000), "The Malmquist Productivity Index and Plant Capacity Utilization", *Scandinavian Journal of Economics* 102, 303-310.

Diewert, W.E. (1974), "Applications of Duality Theory", pp. 106-171 in *Frontiers of Quantitative Economics*, Volume 2, M.D. Intriligator and D.A. Kendrick (eds.), Amsterdam: North-Holland.

Diewert, W.E. (1976), `"Exact and Superlative Index Numbers", *Journal of Econometrics* 4, 115-145.

Diewert, W.E. (1992), "Fisher Ideal Output, Input and Productivity Indexes Revisited", *Journal of Productivity Analysis* 3, 211-248.

Diewert, W.E. (2002), "The Quadratic Approximation Lemma and Decompositions of Superlative Indexes", *Journal of Economic and Social Measurement* 28, 63-88.

Diewert, W.E. (2009), "Cost of Living Indexes and Exact Index Numbers", pp. 207-246 in *Quantifying Consumer Preferences*, Daniel Slottje (ed.), Contributions to Economic Analysis Series, United Kingdom: Emerald Group Publishing.

Diewert, W.E. and D. Lawrence (2005), "The Role of ICT in Australia's Economic Performance", Chapter 3 (pp. 55-94) in *ICT and Australian Productivity: Methodologies and Measurement*, Australian Government, Department of Communications, Information Technology and the Arts, Canberra: Commonwealth of Australia. http://www.econ.ubc.ca/diewert/roleper.pdf

Diewert, W.E., T. Nakajima, A. Nakamura, E. Nakamura and M. Nakamura (2005), "The Definition and Estimation of Returns to Scale with an Application to Japanese Industries", unpublished manuscript.

Diewert, W.E. and K.J. Fox (2008), "On the Estimation of Returns to Scale, Technical Progress and Monopolistic Markups", *Journal of Econometrics* 145, 174-193.

Färe, R. and S. Grosskopf (2004), *New Directions: Efficiency and Productivity*, Boston/London/Dordrecht: Kluwer Academic Publishers.

Färe, R., S. Grosskopf, M. Norris and Z. Zhang (1994), "Productivity Growth, Technical Progress, and Efficiency Change in Industrialized Countries", *American Economic Review* 84, 66-83.

Färe, R., S. Grosskopf and P. Roos (1998), "Malmquist Productivity Indexes: A Survey of Theory and Practice", in R. Färe, S. Grosskopf and R. R. Russell (eds.), *Index Numbers: Essays in Honour of Sten Malmquist*, Boston/London/Dordrecht: Kluwer Academic Publishers.

Färe, R., S. Grosskopf and R.R. Russell (eds.) (1998), *Index Numbers: Essays in Honour of Sten Malmquist*, Boston/London/Dordrecht: Kluwer Academic Publishers.

Fisher, I. (1922), *The Making of Index Numbers*, Houghton Mifflin: Boston.

Fox, K.J. (ed.) (2002), *Efficiency in the Public Sector*, Boston/London/Dordrecht: Kluwer Academic Publishers.

Fox, K.J., R.Q. Grafton, J. Kirkley and D. Squires (2003), "Property Rights in a Fishery: Regulatory Change and Firm Performance", *Journal of Environmental Economics and Management* 46, 156-177.

Grosskopf, S. (2003), "Some Remarks on Productivity and its Decompositions", *Journal of Productivity Analysis* 20, 459-474.

Guo, J-T. (2004), "Increasing Returns, Capital Utilization, and the Effects of Government Spending", *Journal of Economic Dynamics and Control* 28, 1059-1078.

Guo, J.-T. and K.J. Lansing (2002), "Fiscal Policy, Increasing Returns, and Endogenous Fluctuations", *Macroeconomic Dynamics* 6, 633-664.

Hailu, A. and S.T. Veeman, "Environmentally Sensitive Productivity Analysis of the Canadian Pulp and Paper Industry, 1959-1994: An input distance function approach", *Journal of Environmental Economics and Management* 40, 251-274.

Hicks, J.R. (1961), "Measurement of Capital in Relation to the Measurement of Other Economic Aggregates", in F.A. Lutz and D.C. Hague (eds.), *The Theory of Capital*, London: Macmillan.

Hintermaier, T. (2003), "On the Minimum Degree of Returns to Scale in Sunspot Models of the Business Cycle", *Journal of Economic Theory* 110, 400-409.

Hollingsworth, B. (2004), "Non Parametric Efficiency Measurement", *Economic Journal* 114, F307-F311.

Jones, C.I. (2004), "Growth and Ideas", NBER Working Paper 10767, Cambridge, MA: National Bureau of Economic Research.

Jorgenson, D.W. and Z. Griliches (1967), “The Explanation of Productivity Change”, *Review of Economic Studies* 34, 249–283.

Kohli, U. (2003), “GDP Growth Accounting: A National Income Function Approach”, *Review of Income and Wealth* 49, 23-34.

Kruger, J.J. (2003), “The Global Trends of Total Factor Productivity: Evidence from the nonparametric Malmquist Index Approach”, *Oxford Economic Papers* 55, 265-286.

Kumar, S. and R.R. Russell (2002), “Technological Change, Technological Catch-Up, and Capital Deepening: Relative Contributions to Growth and Convergence”, *American Economic Review* 92, 527-548.

Laitner, J. and D. Stolyarov (2004), “Aggregate Returns to Scale and Embodied Technical Change: Theory and Measurement Using Stock Market Data”, *Journal of Monetary Economics* 51, 191-233.

McIntosh, J. (2002), “A Welfare Analysis of Canadian Chartered Bank Mergers”, *Canadian Journal of Economics* 35, 457-475.

Malmquist, S. (1953), “Index Numbers and Indifference Surfaces”, *Trabajos de Estastistica* 4, 209-242.

Moorsteen, R.H. (1961), “On Measuring Productive Potential and Relative Efficiency”, *Quarterly Journal of Economics* 75, 451-467.

Nakajima, T., M. Nakamura and K. Yoshioka (1998), “An Index Number Method for Estimating Scale Economies and Technical Progress Using Time-Series of Cross-Section Data: Sources of Total Factor Productivity Growth for Japanese Manufacturing, 1964—1988”, *Japanese Economic Review* 49, 310-334.

Nin, A., C. Arndt and P.V. Preckel (2003), “Is Agricultural Productivity in Developing Countries Really Shrinking? New Evidence using a Modified Nonparametric Approach”, *Journal of Development Economics* 71, 395-415.

Norman, V.D. and A.J. Venables (2004), “Industrial Clusters: Equilibrium, Welfare and Policy”, *Economica* 71, 543-558.

Ozgen, H. and Y.A. Ozcan (2004), “Longitudinal Analysis of Efficiency in Multiple Output Dialysis Markets”, *Health Care Management Science* 7:4, Special Issue Nov., 252-261.

Shiu, A. (2003), “Multilateral Comparisons of Productivity, Terms-of-Trade and Factor Accumulation” *Review of Income and Wealth* 49, 35-52.

Solow, R.M. (1957), “Technical Change and the Aggregate Production Function”, *Review of Economics and Statistics* 39, 312-320.

Sturm, J-E. and B. Williams (2004), “Foreign Bank Entry, Deregulation and Bank Efficiency: Lessons from the Australian Experience”, *Journal of Banking and Finance* 28, 1775-1799.

Tinbergen, J. (1942), “Zur Theorie der lanfristigen Wirtschaftsentwicklung”, *Weltwirtschaftliches Archiv* 55, 511-549.

Wang, C.J. (2003), “Productivity and Economies of Scale in the Production of Bank Service Value Added”, Federal Reserve Bank of Boston, Working Papers: 03-7.

Weber, W.L. and B. Domazlicky (2001), “Productivity Growth and Pollution in State Manufacturing”, *Review of Economics and Statistics* 83, 195-199.