Productivity Decomposing Model: Theoretical Presentation

Bassim Shebeb*

ملخص

نظرية نمذجة الإنتاجية الكلية ومكوناتها الرئيسية

يعتبر مؤشر الإنتاجية الكلية واحد من أهم مؤشرات الأداء الاقتصادى إضافة إلى أنه يتسم بالشمولية، ومن الممكن تحليله إلى أربعة مكونات رئيسية كمؤشرات للأداء الاقتصادى متمثلة فى: (١) الكفاءة الإنتاجية، (٢) عائد الغلة للحجم، (٣) معدل استغلال الطاقة الإنتاجية، (٤) التغير التكنولوجى. وتكمن أهمية التعرف على المكونات الرئيسية لمؤشر الإنتاجية الكلية فى الحصول على تقديرات تتسم بالدقة والقابلية العالية فى تحليل أسباب التغيرات فى مستويات الإنتاجية المختلفة، وعليه فإن تجاهل نمذجة وقياس هذه المكونات يؤدى فى معظم الأحيان إلى استنتاجات غير دقيقة عن الأسباب الحقيقية وراء التغيرات فى مستوى الإنتاجية. وبناء على ذلك كانت هناك حاجة ملحة لإيجاد نموذج شامل يأخذ بعين الاعتبار أثر التغيرات فى هذه المكونات الرئيسية على قياس وتحليل الإنتاجية الكلية.

وتهدف هذه الدراسة إلى التعرف على النموذج الملائم للاستخدام فى قياس وتحليل الإنتاجية الكلية ومكوناتها الرئيسية وذلك من خلال القيام بمراجعة وعرض الدراسات الحديثة فى هذا المجال. كما تبين الدراسة نظرية شاملة لنمذجة الإنتاجية الكلية وأهم مكوناتها.

* Assistant Professor, Department of Economics and Finance, University of Bahrain. The author would like to thank Professor H. Campbell for the comprehensive feedback and suggestions. The comments of Dr. J. Longmire are also highly appreciated.

1. Introduction

Over the last two decades, applied economists realized that the fundamental issues of isolating the contribution of scale economies and change in capacity utilization and efficiency to productivity growth were remained unsolved. However, as a result of recent developments, productivity growth could be decomposed into several important measures of economic performance, Shebeb (1998). These measures are mainly, technical change, scale economies, productive efficiency, and capacity utilization. The challenge that is undertaken in this paper is to visualize the decomposition of productivity growth and comprehend the underlying theory.

The paper is organised as follows. Section 2 discusses the underlying theory of the measurement of productivity and its associated economic performance indicators. The discussion of this Section includes a brief review of dual cost productivity measurement and its linkage to disembodied technical change and scale economies. Section 3 reviews the traditional and economic measures of capacity utilization and its linkage to productivity growth. In Section 4, the discussion is directed to the frontier-based productivity decomposition modelling. The discussion of this section includes a brief review of productivity efficiency and its linkage to productivity growth. Next, a model for decomposing productivity growth is presented. The Conclusion is presented in the final section.

2. Productivity Measurement: A Dual Cost Approach

Productivity can be defined either by increased output holding the level of inputs unchanged or reduced cost of production holding the level of output unchanged. These definitions can, however, be presented theoretically either by an upward shift of the isoquant or by a downward shift in the average cost function. Thus, the production and/ or cost function can be used to represent the underlying technology and to develop the theoretical linkage between productivity growth and its major components, Diewert (1992).

However, most recent developments in productivity measurement and analysis are based on the convexity and derivative properties of the dual cost function. In the modern approach to productivity measurement, productivity growth is measured in terms of cost saving for given levels of output rather than output-increasing for given levels of inputs. That is, the fundamental concept underlying the cost-based measure of multi-factor productivity growth is that if a given output can be produced with a smaller amount of inputs due to technological improvement, it implies that this level of output may be produced at a lower cost, in real terms.

A cost function may be defined as $C = C(Q, P, t)$, where C is the total cost, Q is the output level, P is a vector of the input prices, and t is a time trend employed as a proxy for technology. This cost function is assumed to be the lowest cost for a given level of output Q, given input prices and technology. This cost function needs to satisfy the regularity conditions for a well-behaved cost function. It follows that the change in cost over time holding output and input prices unchanged reflects the change in productivity, Shephard (1953, 1970) and McFadden (1966, 1978).

However, the observed change in overall productivity (MFP) can be a result of various economic interactions in the production process, including technical change, scale economies, and changes in capacity utilization and inefficiency. If any of these major economic aspects of the production process is ignored, the resulting estimates of MFP are likely to have measurement bias. It follows that a full structural model is needed to decompose productivity growth into its major components. In what follows a full structural model is constructed.

3. Dual Cost Measurement of Capacity Utilization and Its Linkage to Productivity Growth

It was Cassels (1937) who first recognized that capacity utilization is a reflection of scarcity or "fixity" of the production factors that are available to a firm. Input scarcity can be seen as a short-run constraint to the economic optimisation of a firm. These short-run constraints would make the short-run minimum cost level of output differ from that of the long run. An economic theory-based measure of capacity utilization which takes account of these inputs fixity constraints and measures the optimal level of output given these constraints is developed to obtain a highly interpretable measure of capacity utilization. The determination of the optimal level of output under input fixity was originally presented by Klein (1960) following the clear distinction between excess capacity in the short-run and that in the long-run which was made by Cassels (1937). The capacity output is defined to be that level of output at the tangency point of short- and long-run average cost curves.

Berndt and Morrison (1981 p52) conclude their study with "we hope that applied researchers in the future will devote greater attention and care to the economic theory underlying the concept of capacity.... Which can then be interpreted more clearly." Thus, an estimate of capacity utilization that is based on economic theory is needed to provide more reliable and rigorous dynamic explanations of economic performance.

Referring to Figure 1 the capacity output might be measured by Q^* If the capacity level of production is equal to the output level in the long-run equilibrium for a given set of input constraints, full capacity utilization will exist. It implies that both capacity output and the observed output are equal. However, when the demand for output is less, Q^{ti}, or greater Q^0, than the output level in long-run equilibrium the capacity measure would show that capacity is underutilised, $Q^U/Q^*<1$, or over-utilized, $Q^0/Q^*>1$.

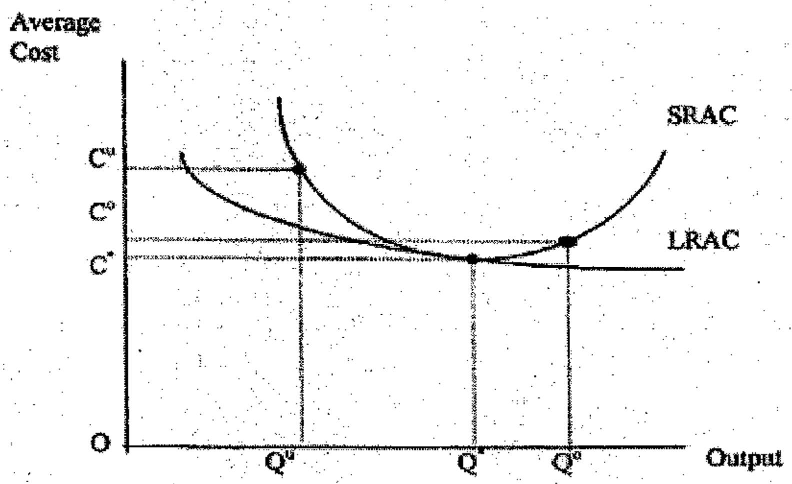

Figure 1. Capacity Output Under Inputs Fixity Allowing For Non-constant Returns to Scale.

This approach takes explicitly into account the fixity of different inputs that may occur in the short-run production process. It also determines the firm's optimal responses under the fixity of these inputs. The main economic aspect underlying the cost-based measure of capacity utilization is the degree of fixity of the scarce production factors. Thus, input fixity is the key factor that causes capacity not to be fully utilised in the short run. This implies that a measure of capacity utilization can be based on short-run specification of cost structures which reflect underlying production relationships, Morrison (1985, 1988a, 1988b) and Segerson and Squires (1990).

Estimating this measure of capacity utilization is not straightforward, since it requires an estimation of the short-run and long-run average costs curves. It also needs a relatively flexible form of cost function. However, the recent developments in dual cost theory have established a well-defined method for empirical estimation. This method could explicitly restrict the cost function for short-run input fixity.

3.1 The Linkage between Productivity Growth and Capacity Utilization

The linkage between capacity utilization and productivity growth could be one of the most important aspects involved in the interpretation of the productivity measure over the short-run. In the above discussion, the concept of capacity utilization and its cost-based measure was presented. It was indicated that the level of capacity utilization plays an important role in adjusting multi-factor productivity change for input fixity in the short-run. In this section a diagrammatic exposition of the relationship between productivity growth and capacity utilisation with non-constant returns to scale is presented.

Take an industry made up of identical firms whose input-output coefficients do not vary across firms at any level of output. Cost and production theory suggests that a reasonable representation of the industry, presumed to be realising increasing returns to scale. As in Figure 2, the downward slope of the long-run average cost curve (LRAC) implies that returns to scale are increasing, but at a decreasing rate with additional output. The short-run average cost (SRAC) curve is tangential to the LRAC at Y.

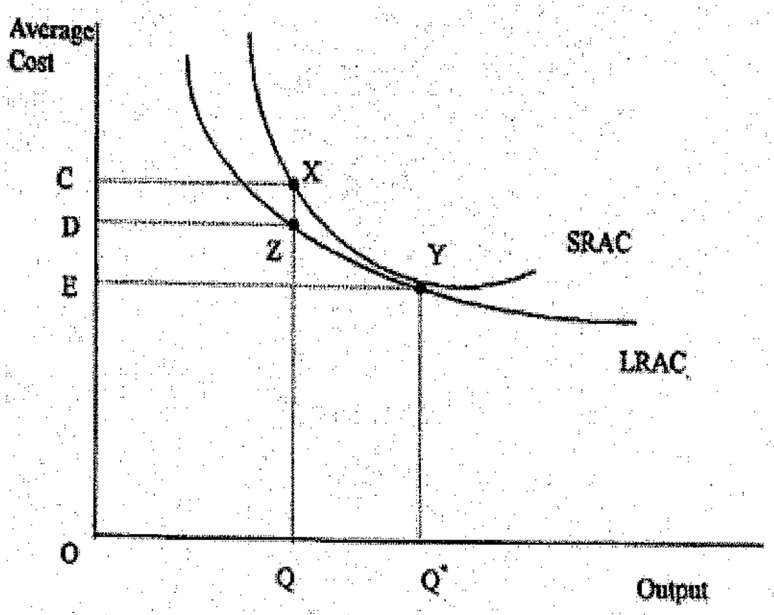

Figure 2. Productivity Growth Allowing for Non Constant Returns to Scale, Capacity Utilisation, and No Technical Change.

Referring to Figure 2 above, when the industry is operating at Q^*, it is fully utilising its capacity. At this level of output, short-run cost equals long-run cost and no scope exists for further cost reduction by changing the capacity or the capacity utilization of the industry. In other words, at Y the industry has the existing capacity that matches the desired level of output. More commonly, however, industries operate at levels of output other than Q^*.

Recall that in Figure 2 the level of technology in the industry is assumed to be fixed. Without technological change, any measured productivity change would be attributable to scale and capacity utilization effects. Suppose the industry is operating at level of output Q, and moves from Q to Q*. Consequently, the short-run cost per unit will decline by CE and this would translate into productivity gain that is proportional to the overall cost reduction. Since there is no technical change, two effects underlie this decline in costs associated with the movement from X to Y: (1) the cost reduction that is associated with better utilising plant capacity, CD; and (2) the cost reduction that is associated with realising returns to scale, DE. That is, productivity growth may be realised with *no* technological change.

The simple analysis of Figure 2 highlights the point that measured productivity change, based on measured cost reduction through time, might be the result of scale economies and/or changing capacity and capacity utilization in the industry rather than technological change. Now introduce technological change, as in Figure 3. In this case the SRAC curves and LRAC curves shift downward as a result of the change in technology to SRAC* and LRAC* respectively.

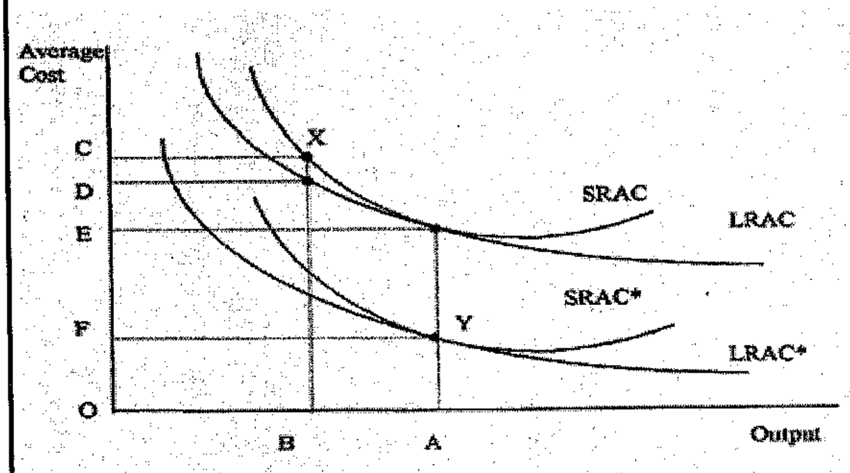

Figure 3. Productivity Growth: Allowing for Non-Constant Returns to Scale, Capacity Utilisation, and Technical Change.

If the industry moves from X to Y in Figure 3 it might be a result of three sources of potential cost reduction: (1) the technological change effect, EF; (2) the returns to scale effect, DE; and (3) the capacity utilization effect, CD. The overall gain in productivity that would be measured by the overall cost reduction CF could now be attributed to these three sources.

The main point of the above analysis was to illustrate that measured productivity growth can be decomposed into three components; technical change, scale economies, and change in capacity utilization. Also very important is the fact that failing to account for returns to scale and capacity utilization effects will tend to bias measured rates of productivity growth.

4. Productivity Growth and Productive Efficiency

So far the discussion on productivity growth and its decomposition has been based on the assumption that the production process exhibits productive (technical and allocative) efficiency. That is, the measurement of productivity presented in the previous section has ignored the possible contribution of the changes in productive efficiency or its components to productivity growth. In this Section, a change in productive (cost) efficiency as a component of productivity change is considered. Isolating the cost efficiency component and measuring its effect on MFP growth is the purpose of this Section.

Most of the productivity and production studies assume implicitly, and sometimes explicitly, that the production process is efficient. This implies that the producer is efficient- that s/he achieves her/his economic objectives. In a productivity sense, this assumption implies that producers always operate at the production (cost) frontier and any change in productivity is the result of a shift in the frontier however, the inefficient producers are operating below (above) the production (cost) frontier. In this case, the change in productivity should not be referred to as a shift in the frontier, but should be considered as a movement towards the frontier. Thus, to avoid any misinterpretation of productivity estimates, the impact of inefficiency needs to be identified in the productivity measurement model, Fried et al. (1993), and Nishimizu and Page (1982).

To make the above argument more clear a graphical presentation follows. This presentation is conducted based on a simple production/cost structure: that is constant returns to scale and full capacity utilization are assumed, Grosskopf (1993). Figure 4 shows the long-run and short-run average cost curves of a technically efficient producer. There are two time periods t and t+1 with two observed average cost levels (Q^*, C_t) and (Q^*, C_{t+1}) respectively. Neither of these cost levels are located at the cost frontier itself.

Productivity growth or technical change can be measured for an efficient producer by the reduction in cost of producing a given level of output. That is $\partial C(.)/\partial t<0$ when there is a productivity improvement. Now suppose that the producer is not cost efficient. It implies that it produces a given level of output with higher cost than the potential minimum average cost. That is, the cost of production will be located above the cost frontier in the two time periods, t and t+1. The non-frontier productivity measurement model, which ignores efficiency as a component of productivity, would consider the observed cost as equivalent to the minimum cost, at the frontier. Figure 4 shows the difference between the observed and the minimum possible cost (cost-efficient) in two-time periods t and t+1.

According to our earlier analysis, a non-frontier measure of productivity would consider the reduction in cost from C_t to C_{t+1} as a measure of technical change. However, the actual measure of technical change is the reduction in C^*_t to C^*_{t+1}. it follows that the difference between the two measures will be due to the existence of the cost inefficiency, $(C_{t+1} - C_t) \neq (C^*_{t+1} - C^*_t)$. Therefore, to obtain the estimate of technical change it is necessary to correct the observed cost levels to bring them down to the cost frontier in both periods. Once the observed cost levels have been corrected for the existence of cost inefficiency, estimates of the productivity growth and its major components can be obtained.

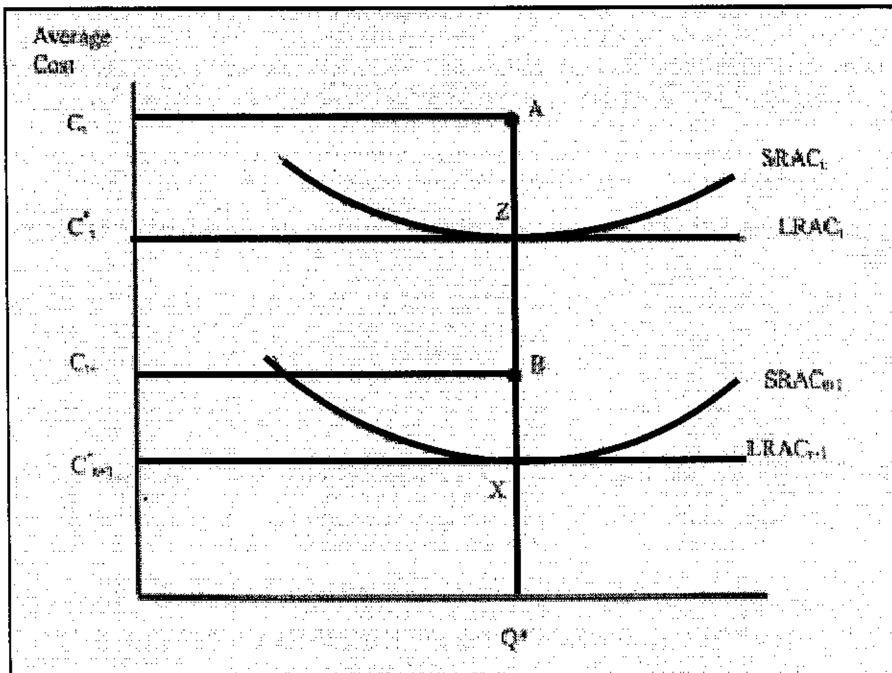

Figure 4 . Cost-Based Measure of Productivity Growth Allowing for Cost inefficiency.

The corrected measures of the observed cost levels may be obtained by defining the minimum cost level as; $C^*_t = D_t(C_t, Q^*) . C_t$, and $C^*_{t+1} = D_{t+1}(C_{t+1}, Q^*) . C_{t+1}$, Where $D_t(C_t, Q^*)$ and $D_{t+1}(C_{t+1}, Q^*)$ are distance functions in time t and t+1 respectively, Grosskopf (1993).

Thus, observed cost saving over the two periods t and t+1, $(C_{t+1} - C_t)$ can be seen as a result of cost-saving due technical change, $(C^*_{t+1} - C^*_t)$, and that due to change in cost inefficiency which may be measured as $(D_{t+1} - D_t)$. It follows that this frontier-based productivity growth is decomposed into two main parts; technical change and the change in cost inefficiency, given that full capacity utilization and constant returns to scale are assumed.

To conclude this simple graphical presentation, a measure of multi-factor productivity growth based on a non-frontier model could not be interpreted as a gain due to shift in cost/ production frontiers unless it is assumed that there is no change in cost efficiency over time. This type of decomposition of MFP growth has quite important policy implications.

5. Productivity Decomposing Model

To summarize the theoretical linkages between productivity growth and its major decompositions (technical change, scale economies, cost inefficiency, and capacity utilization) a comprehensive diagrammatic exposition is now presented. Once again, take an industry made up of identical firms whose input-output coefficients do not vary across firms at any level of output. Cost and production theory suggest that a reasonable representation of the representative firm in the industry, presumed to be realising increasing returns to scale, is as in Figure 5. The downward slope on the long-run average cost curve (LRAC) implies that returns to scale are increasing, but at a decreasing rate, with additional output.

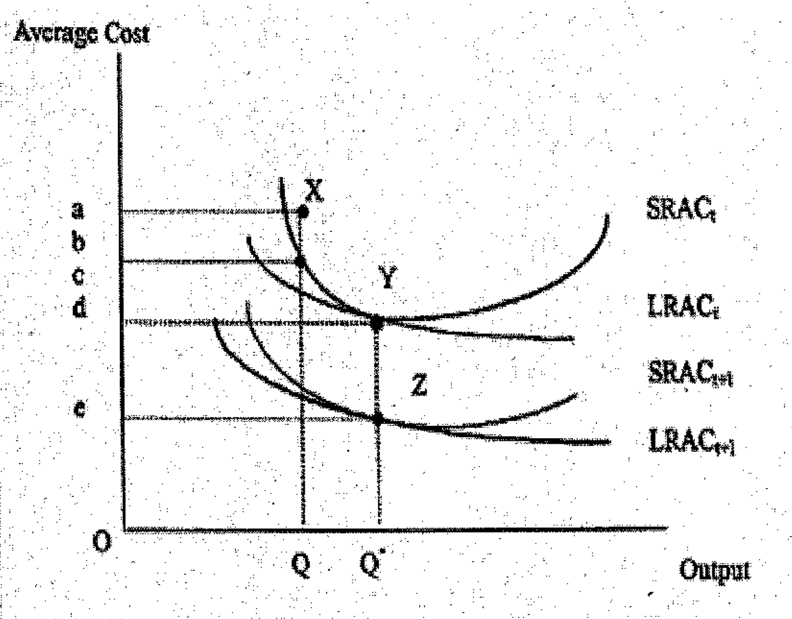

Figure 5. Productivity Decomposing Model: Allowing for Cost Inefficiency, Technological Change, Capacity Utilisation, and Non-Constant Returns to Scale.

The short-run average cost, $SRAC_t$ curve is tangential to the $LRAC_t$ at Y *(Hence: disregard $SRAC_{t+1}$ and $LRAC_{t+1}$ curves in Figure 5 for the time being)*. When the firm is operating at capacity output level Q^*, it is said to be fully utilising its capacity (in an economic sense) and it is cost efficient. At this level of output, Q^*, the short-run average cost is equal to that of the long-run, hence, no scope exists for further cost reduction by changing the capacity or the capacity utilization of the firm. More commonly, however, firms operate at levels of output other than

Q^*. For example, if it is operating at output level Q, and some cost inefficiency does exist, the average cost in this case would be equal to Oa which exceeds the SRAC and $LRAC_t$ by the amount ab and bc per unit of output, respectively. The ab portion is attributed to the cost inefficiency while the (ac - ab) = bc portion can be ascribed to capacity under-utilization in the short run.

Assume that the firm was operating at level of output Q with an average cost of Oa, and that it then moves from X to Y. Short-run cost per unit will decline by ad and this would be translated into a gain in multi-factor productivity. Three possible effects underlie the decline in cost associated with the movement from Q to Q^* These are as follows. (1) the efficiency gain (cost reduction, ab) associated with operating at the cost frontier; (2) the cost reduction associated with realising the long-run average cost bc and (3) the cost reduction due to the realization of scale economies cd Noting that, if the move from Q to Q^* involves no change in plant capacity, the full productivity gain, ad would be attributed to better capacity utilization and full cost efficiency only. On the other hand, if the move from Q to Q^* involves expanding the plant capacity by QQ^* then cd will represent the cost savings from the scale effect, as mentioned above. It follows that the remainder of ad would represent the cost savings from better utilising capacity and cost efficiency. This analysis highlights the fact that productivity change, based on measured cost reduction, might be the result of changing the scale of the operation, capacity utilization and cost efficiency in the firm rather than technological change. Thus, without technological change, any productivity gain should be attributable to scale and capacity utilization and cost efficiency components.

Now let us introduce technological change with the result that the $SRAC_t$ and $LRAC_t$ curves can be shifted downward as a consequence of technological improvement as depicted by $SRAC_{t+1}$ and $LRAC_{t+1}$ respectively.

Take it that the firm moves from X to Z in Figure 5. Thus, in this case, there are four sources of potential cost reduction that can be analysed. These sources are associated with: (1) cost efficiency, ab; (2) the capacity utilization effect, bc; (3) the realization of scale effect cd; and (4) the technological change, de. The overall gain in productivity would thus be measured by the overall cost reduction ae, which now could be ascribed to these four sources. To generalise the simplified analysis presented above we need to consider all combinations of levels of Q^* and Q, including moving from one point where capacity is/ is not fully utilised and/or from points where no gain is associated with cost efficiency. That was not illustrated diagrammatically for ease of exposition.

6. Concluding Remarks

Productivity growth is a composed measure of a number of economic behaviors that are important pieces of the overall economic performance puzzle. Identifying and measuring these components of overall productivity helps to provide a more accurate and interpretable measure of economic performance. That is, the observed change in overall productivity (MFP) could be a result of various economic interactions in the production process, including technical change, scale economies, and changes in capacity utilization and inefficiency. It follows that if any of these major economic aspects of the production process is ignored, the resulting estimates of MFP are likely to have measurement bias. In this paper, a hill structural model is constructed to take into account the contribution of these major components to the overall productivity growth measure.

References

Berndt, E.R. & M. Fuss 1989, "Economic Capacity Utilization and Productivity Measurement for Multi-product Firms with Multiple Quasi-Fixed Inputs", *Working Paper,* no.2932, National Bureau of Economic Research, April, pages 36.

Berndt, E.R. & D. Hesse 1986, "Measuring and Assessing Capacity Utilization in the Manufacturing Sectors of Nine OECD Countries", *European Economic Review,* vol. 30, pp. 961-989.

Berndt, E.R. & C. J. Morrison 1981, "Capacity Utilization Measures: Underlying Economic Theory and Alternative Approach", American Economic Review, vol. 71, No. 2, pp. 48-57.

Cassels, J. M. 1937, "Excess Capacity and Monopolistic Competition", Quarterly Journal of Economics, vol. 51, pp. 426-443.

Fried H. O., C.A.K. Lovell, & 5.5. Schmidt eds. 1993, The Measurement of Productive Efficiency: Techniques and Applications, Oxford: Oxford University Press.

Grosskopi, S. 1993, "Efficiency and Productivity", in H.O. Fried, C.A.K Lovell, & S.S. Schmidt eds., The Measurement of Productive Efficiency: Techniques and Applications, Oxford: Oxford University Press.

Hickman 1964, "On a New Method of Capacity Estimation", Journal of the American Statistical Association, vol. 59, No. 306, pp. 529-549.

Klein, L.R. 1960, "Some Theoretical Issues in the Measurement of Capacity", Econometrics, April, vol. 28, pp. 27186.

Kopp, R.J. 1981, "The Measurement of Productive Efficiency: A Reconsideration", Quarterly Journal of Economics, 96, Aug., 477-504.

Kopp, R.J. & W. Diewert 1982, "The Decomposition of Frontier Cost Function Deviations into Measures of Technical and Allocative Efficiency", Journal of Econometrics, vol 19, No.2/3, pp. 319-332.

McFadden, D. 1966, "Cost, Revenue and Profit Functions: A Cursory Review Working Paper, no.86, Institute for Business and Economic Research, University of California.

McFadden, D. 1978, "Cost, Revenue and Profit Functions: A cursory Review", in M. Fuss & D. McFadden eds., Production Economics: A Dual Approach to Theory and Applications, vol.1, Amsterdam: North-Holland.

Morrison, C.J. 1985, "Primal and Dual Capacity Utilization: An Application to Productivity Measurement in the U.S. Automobile Industry", Journal of Business and Economic Statistics, vol 3, No.4, October, pp. 31224.

Morrison, C.J. 1983a, "Capacity Utilization and Productivity Measurement: An Application to the U.S Automobile Industry", in A. Dogramaci ed. Studies in Productivity Analysis, Boston: Kluwer Nijhoff Press, pp. 163-193.

Morrison, C.J. 1988b, "Quasi-Fixed Inputs in U.S. and Japanese Manufacturing:
A Generalized Leontief Restricted Cost Function Approach", Review of Economics and Statistics, vol. 70, No. 2, May, pp. 275-287.

Nadiri, I.M. 1970, "Some Approaches to the Theory and Measurement of Total Factor Productivity: A Survey", Journal of Economic Literature, vol. 8, No. 4, pp. 1137-1177.

Nishimizu, M. & J.M. Page 1982, "Total Factor Productivity Growth, Technological progress and Technical Efficiency Change: Dimensions of Productivity Change in Yugoslavia 1965-78", Economic Journal No. 92, pp. 920-936.

Segerson, K. & D. Squires 1990, "On the Measurement of Economic Capacity Utilization for Multi-Product Industries", Journal of Econometrics, vol. 44, No. 3, pp. 347-361.

Shebeb, B. 1998, "Measuring Sources of Multi-Factor Productivity Change Using Stochastic Frontier and Non-Frontier Approaches: An Application to the Australian Mining Industries", Ph.D. thesis, The University of Queensland.

Shephard, R.W. 1953, Cost and Production Functions, Princeton: Princeton University Press.

Shephard, R.W. 1970, Theory of Cost and Production Functions; Princeton: Princeton University Press.