Aggregation of Productivity Indices: The Allocative Efficiency Correction

Thijs ten Raa[1]
Tilburg University
5000 LE Tilburg
the Netherlands
tenRaa@UvT.nl

January 9, 2005

Abstract. Aggregation of productivity indices is known to be consistent under very restrictive conditions only. This paper quantifies the aggregation inconsistency.

Keywords: Aggregation, productivity, allocative efficiency

JEL codes: O47, C43

[1] I am grateful to three anonymous referees and editor Russell for suggestions that improved the paper.

1. Introduction

Firms' productivity indices do not sum to the industry productivity index, except when production is linear in the sense that marginal rates of substitution and marginal rates of transformation are constant and these constants are common to the firms (Blackorby and Russell, 1999). The trouble is that industry productivity is influenced not only by the performance of firms, but also by the allocation of resources between the firms. In an attempt to salvage the aggregation of productivity, Färe and Primont (2003) show that if all firms are allocatively efficient and their technologies admit time-invariant quadratic approximations, then the productivity indices can be aggregated. Unfortunately, the Färe-Primont conditions are also prohibitively restrictive. To me the bottom line seems to be that the determination of industry productivity requires not only the aggregation of firm productivities, but also the inclusion of some allocative efficiency terms. Indeed, in a recent growth accounting study, Jorgenson, Ho, and Stiroh (2003) aggregate industry productivities to total productivity growth and capture allocative efficiency changes in their formula (53). Johansen (1972) offers a simple but powerful framework that I will use for a back-of-the-envelop theory that encompasses the different strands of the productivity aggregation literature and quantifies the aggregation inconsistency.

Unfortunately, the literature is loaded with formulas. Part of the blame can be put on the mix of conceptual and approximation issues. A conceptual issue is the definition of productivity. Roughly speaking, it is the output/input ratio, but input and output are multi-dimensional objects. An approximation issue is at which level to meter rates of changes. It makes a difference if one approximates the time derivative of an input or the time derivative of the logarithm of the input. Respective measures are the Fisher ideal and the Törnqvist indices. This difference dissolves in the continuous time framework, which I henceforth use for analytical convenience. I also simplify the concept of a productivity indicator. Färe and Primont (2003) use the Luenberger indicator, which is based on the distance to the frontier along some direction in commodity space; they remain silent about the choice of direction. I suggest that, at least for single-output industries, the appropriate Luenberger direction is along output. In this case the production structure is quasi-linear in output. Now Woertman and ten Raa (2004) argue that for quasi-linear functions the direction is determined by the linear commodity component, as in that case Luenberger's measure is equal to both the compensating and the equivalent variations. In the context of consumption theory this is the direction of the numerair commodity and in the context of production theory it is the direction of output. This observation reduces the distance function to the output gap and, as we shall see, the derived productivity indicator to the Solow residual. It makes the analysis so crisp that the extension to multi-output industries becomes obvious.

2. Indices

Let me introduce the formalities. Single-output firm k maps input vector $x_k(t)$ in output scalar $y_k(t) \leq F_k(x_k(t), t)$, where $F_k(\bullet, t)$ is its production function at time t. (Parameter t shifts the production function, or what is called technical change.) In this one-dimensional output case Luenberger's output based distance function is given by

$$D_k(x_k(t), y_k(t), t) = F_k(x_k(t), t) - y_k(t) \tag{2.1}$$

and measures the *output gap* (Färe and Grosskopf, 2004, p. 107). In general, even without the quasi-linear structure of (2.1), the distance function measures *in*efficiency. *Efficiency change* is therefore defined by *minus* the change in the distance function:

$$EC_k = -\frac{dD_k}{dt} \tag{2.2}$$

The distance to the frontier may grow without any change in inputs or outputs, simply because the frontier shifts out. This is called *technical change*. It is defined by the partial derivative of the distance function with respect to time:

$$TC_k = \frac{\partial D_k}{\partial t} \tag{2.3}$$

The sum of efficiency change and technical change defines *productivity change*:

$$PC_k = EC_k + TC_k \tag{2.4}$$

Application of the chain rule to (2.2) and addition of (2.3) transforms (2.4) into

$$PC_k = -\frac{\partial D_k}{\partial x_k}\dot{x}_k(t) - \frac{\partial D_k}{\partial y_k}\dot{y}_k(t) \tag{2.5}$$

Here, as usual, a dot denotes the derivative of a function of time and the derivatives with respect to the input and output vectors are row vectors (so that the terms on the right hand side are inner products which can be denoted without dot). The discrete time approximation of (2.5) is what Chambers, Färe, and Grosskopf (1996) call the Luenberger productivity index. That index is the point of departure of Färe and Primont's (2003) aggregation analysis. In case of the quasi-linear structure of (2.1), expression (2.5) simplifies quite dramatically into

$$PC_k = \dot{y}_k(t) - \frac{\partial F_k}{\partial x}(x_k(t), t)\dot{x}_k(t) \tag{2.6}$$

In other words, productivity change is equal to firm *k*'s *Solow residual* between its output change and input changes, where the latter are weighted by their marginal product values.

3. Aggregation

The more standard Solow residual is at the macro level, or, in the context of the present literature, the industry level. For this I need the *industry* distance function or output gap. Following Johansen (1972), potential industry output is a function of *total input*, $x(t)$ = $\sum x_k(t)$, determined by the following constrained maximization problem:

$$\max \sum F_k(\xi_k, t) : \sum \xi_k = x(t) \quad (3.1)$$

Since the optimal allocation depends on time (through the constraint and the objective function), let me denote it by ξ_k = $x_k{}^*(t)$. In my view the (negative) aggregation results of Blackorby and Russell (1999) and Färe and Primont (2003) reflect the condition that attainment of the optimal industry output requires not only a push of the firms to their respective frontiers, from $y_k(t)$ to $F_k(\cdot, t)$, but also a reallocation of resources between them, that is from $x_k(t)$ to $x_k{}^*(t)$. The benefit of the latter reallocation is simply missed when firm efficiency indices are aggregated, without correction. The missing element is the potential allocative efficiency gain; it will be derived next.

It is important to realize that potential industry output does *not* depend on the allocation of the actual inputs, ($x_k(t)$), and *not even* on the allocation of the optimal inputs, ($x_k{}^*(t)$). Potential industry output depends *only* on *total* input, $x(t)$ = $\sum x_k(t)$. True, the solution to (3.1) can be written $\sum F_k(x_k{}^*(t), t)$, but the $x_k{}^*(t)$'s are functions of $x(t)$. In short, program (3.1) maps *one* input vector, $x(t)$, in potential output. Symbol F denotes the mapping.

Potential industry output is $F(x(t), t)$ and, therefore, the industry output gap of (2.1) becomes

$$D(x(t), y(t), t) = F(x(t), t) - y(t) \quad (3.2)$$

where the last term is defined by $y(t)$ = $\sum y_k(t)$. The productivity analysis of the firm can now be applied to the industry. The data are now inputs $x(t)$, outputs $y(t)$, and production function $F(\bullet, t)$. Subjection to (2.6) yields the following expression for industry productivity change:

$$PC = \dot{y}(t) - \frac{\partial F}{\partial x}(x(t), t)\,\dot{x}(t) \tag{3.3}$$

The "aggregation problem" consists of interrelating the micro- and macro-productivity changes, (2.6) and (3.3), respectively. This boils down to an analysis of the industry production function, F, which is the solution to (3.1).

Denote the Lagrange multipliers of the (vector) constraint in (3.1) by (row) vector w. Since Lagrange multipliers measure the sensitivity of the objective function, F, with respect to the bounds in the constraints, $x(t)$, I obtain

$$w(t) = \frac{\partial F}{\partial x}(x(t), t) \tag{3.4}$$

Now the first order condition of (3.1) with respect to ξ_k reads, in the optimum,

$$\frac{\partial F_k}{\partial x}(x_k{*}(t), t) = w(t) \tag{3.5}$$

This is the well-known result that efficiency implies the equalization of marginal productivities. Substitution of (3.5) in (3.4) and subsequently in (3.3) yields

$$PC = \sum [\dot{y}_k(t) - \frac{\partial F_k}{\partial x}(x_k{*}(t), t)\,\dot{x}_k(t)] \tag{3.6}$$

Comparison of this result with (2.6) shows that aggregation is perfect, in the sense that PC_k sum to PC, if

$$\frac{\partial F_k}{\partial x}(x_k{*}(t), t) = \frac{\partial F_k}{\partial x}(x_k(t), t) \tag{3.7}$$

This condition is indeed fulfilled if marginal productivities are constant, an observation that confirms sufficiency part of the result of Blackorby and Russell (1999). The condition is also fulfilled if the mixes of the observed input vectors are right, i.e. if the observed inputs $x_k(t)$ are collinear with the optimal $x_k{*}(t)$, and returns to scale are constant, an observation that confirms the result of Färe and

Primont (2003). If none these conditions are fulfilled, a correction must be made. In fact, the connection between (2.6) and (3.6) is:

$$PC = \sum PC_k + \sum [\frac{\partial F_k}{\partial x}(x_k(t),t) - \frac{\partial F_k}{\partial x}(x_k*(t),t)]\dot{x}_k(t) \qquad (3.8)$$

It is interesting that the correction consists of a sum of terms, one for each firm. For each firm the correction measures the excess marginal productivities (over and above the competitive, economy-wide ones), weighted by the changes in inputs. The difference in brackets is the excess rate of return, or the difference between the private and social values of inputs.

It is not difficult to understand the correction expression. Suppose firm *k* is under endowed with input 1. Then input 1 is relatively scarce at firm 1, hence will carry a high marginal product or supernormal private value. But the latter is used as a weight in the Solow residual of firm *k*, where the input change contributes *negatively*. In short, the scarcity of input 1 causes a downward bias in the Solow residual of firm *k* when the private value weight is used instead of the social value. The positive correction term (the excess rate of return times the change in the input at firm *k*) offsets the bias.

The aggregation bias of productivity changes can go either way. In terms of efficiency *levels*, however, it goes one way, a fact that is exceedingly simple to demonstrate. The solution to (3.1) exceeds the value without reallocations:

$$F(x(t),t) \geq \sum F_k(x(t),t) \qquad (3.9)$$

In view of (2.1) and (3.2) it follows that

$$D(\sum x_k(t), \sum y_k(t),t) \geq \sum D_k(x_k(t),y_k(t),t) \qquad (3.10)$$

This inequality, noted by Färe and Grosskopf (2004, p. 108), states that industry inefficiency exceeds aggregate firms' inefficiency. The difference is the allocative inefficiency. The salient contribution of this paper is to *quantify* the aggregation inconsistency: the second term on the right hand side of formula (3.8).

4. Generalizations

I have derived this result for quasi-linear output gaps, $F(x(t),t) - y(t) \geq 0$. The extension to general production structures, $F(x(t), -y(t), t) \geq 0$, is obvious. If I redefine $(x(t), -y(t))$ as net input vector $x(t)$, I may simply drop $\dot{y}_k(t)$ from productivity changes (2.6) and (3.3). Since (3.8) does not display output $y(t)$, decomposition formula (3.8) remains valid for general *net* input vectors $x(t)$!

The formula does not even seem to depend on the direction in commodity space along which Luenberger's distance function is measured. It is implicit, however. The appropriate direction is given by the gradient of the general industry production structure. However, in this general setting the optimal input allocation, where the gradient is to be evaluated, depends on the desired mixed of outputs and for the latter we need a criterion. If the industry is a price-taker (e.g., exposed to world competition), then that price vector (e.g., the terms of trade) specifies the Luenberger direction. In general, the direction should be determined by the perfectly competitive equilibrium prices.

Strictly speaking, the program defining the industry production function should feature nonnegativity constraints. The incorporation of these constraints is a straightforward application of the Kuhn-Tucker conditions. First order condition (3.5) is replaced by the following complementary slackness condition:

$$\frac{\partial F_k}{\partial x}(x_k*(t),t) \leq w(t), [w(t) - \frac{\partial F_k}{\partial x}(x_k*(t),t)]x*(t) = 0 \qquad (4.1)$$

In view of equations (3.3) and (3.4) the optimal marginal productivity in formula (3.8) must be replaced by $w(t)$:

$$PC = \sum PC_k + \sum [\frac{\partial F_k}{\partial x}(x_k(t), t) - w(t)]\dot{x}_k(t) \qquad (4.2)$$

Formula (4.2) tells us to subtract the *industry* marginal productivity from the firm's marginal productivity. If there are no nonnegativity constraints, the envelope theorem, see equation (3.5), equates industry marginal productivity with optimal firm's marginal productivity and we are back in the ball game of (3.8).

The modification is quite intuitive. Imagine a single input-single output industry with two firms, $y_1 \leq x_1 \geq 0$ and $y_2 \leq 2x_2 \geq 0$. Then the industry production function is given by $y \leq 2x \geq 0$. Suppose $x_1(t) = x_2(t) = t$, $y_1 = t$ and $y_2 = 2t$. Both firms remain on their efficiency frontiers and, therefore, have zero efficiency change. Technical change is zero too, as the production functions do not change. Hence $PC_1 = PC_2 = 0$. Now look at the industry. Technical change is zero. There is efficiency change though. The distance function, (3.2), reads $2x - y$. As a function of time it reads $2(t + t) - (t + 2t) = t$. By the industry version of efficiency change expression (2.2) the latter becomes –1. Indeed, the growth of the inferior firm, 1, reduces efficiency. The allocation is wrong. All input should go to firm 2. The allocative efficiency changes, the second sum in (4.2), capture this. By equation (3.4) the shadow price is 2. The allocative efficiency change imputed to the first firm is $[1 - 2]1 = -1$. The allocative efficiency term efficiency term of the second firm is zero. If we would not have Kuhn-Tucker modified equation (3.8), but compared the actual marginal productivity to the optimal marginal productivity of firm 1, we would have missed the allocative efficiency term, simply because its marginal productivity is constant. The upshot is that firm 1's productivity is so low that it is irrelevant as a benchmark for efficiency change measurement.

An alternative procedure to incorporate nonnegativity constraints is to absorb them in the general production structure, $F(x(t), -y(t), t) \geq 0$. Then the just given generalization of (3.8) can be applied to net input structures. This approach, however, requires the use of subgradients instead of partial derivatives, and is not pursued here.

5. Concluding remarks

Aggregate productivity is the sum of firm productivities *and* firm allocative efficiency changes. A firm's allocative efficiency change is measured by its excess marginal productivities (over and above the competitive economy wide ones), weighted by input changes.

The allocative efficiency correction term requires knowledge of the production function for each firm at each point in time, or, at the very least, the marginal products at both the observed inputs and at the optimally allocated inputs for each firm. Diewert (1992) largely dismisses the possibility of estimating these production functions. However, the logic of the aggregation of efficiency extends to the industries that make an economy; see Jorgenson, Ho, and Stiroh (2003).

References

Blackorby, C. and R.R. Russell, “Aggregation of Efficiency Indices,” *Journal of Productivity Analysis* 12, 1, 5-20 (1999)

Chambers, R., R. Färe and S. Grosskopf, “Productivity Growth in APEC Countries,” *Pacific Economic Review* 1, 3, 181-90 (1996)

Diewert, W.E., “The Measurement of Productivity,” *Bulletin of Economic Research* 44, 3, 163-98 (1992)

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

Färe, R. and D. Primont, “Luenberger Productivity Indicators: Aggregation Across Firms,” *Journal of Productivity Analysis* 20, 3, 425-435 (2003)

Johansen, L., *Production Functions*, North-Holland, Amsterdam (1972)

Jorgenson, D., M. Ho, and K. Stiroh, “Growth of US Industries and Investments in Information Technology and Higher Education,” *Economic Systems Research* 15, 3, 279-325 (2003)

Woertman, W. and Th. ten Raa, “Benefit and Expenditure Functions,” draft, Tilburg University (2004)