Session Number: POSTER PAPER SESSION

Paper Prepared for the 29th General Conference of
The International Association for Research in Income and Wealth

Joensuu, Finland, August 20 – 26, 2006

The Theory of Benchmarking and the Measurement of Industrial Organization

Thijs ten Raa

For additional information please contact:

Author Name(s) : Thijs ten Raa
Author Address(es) : Department of Econometrics and Operations Research, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
Author E-Mail(s) : tenRaa@UvT.nl
Author FAX(es) : +31-13-4663280
Author Telephone(s) : +31-13-4662365

This paper is posted on the following websites: http://www.iariw.org

The Theory of Benchmarking and the Measurement of Industrial Organization

Thijs ten Raa*

June 2006

Abstract

If more productive firms grow relatively fast, an industry performs better, even when no firm exhibits technical or efficiency change. In other words, the two well-known sources of productivity growth–technology and efficiency–can be augmented by a third one, namely the industrial organization effect. In this paper the efficiency of an industrial organization and its contribution to performance are measured by benchmarking all firms on the industry. More precisely, efficiency is measured by the proximity between a firm and the best practices. Aggregation of firm efficiencies is imperfect. The bias is used to measure the efficiency of the industrial organization. In benchmarking, change transmitted by a firm represents productivity growth and change transmitted by the best practices represents technical change. Although I use a nonparametric framework, which requires only input and output information, duality analysis reveals the Solow residual. In discrete time Malmquist indices capture the measurement of the industrial organization effect, efficiency changes, and technical change. The industrial organization of Japanese banking is analyzed. The dynamic industrial organization effect of entry and exit can be accommodated.

JEL Classification Numbers: L10, D24, O47

Keywords: Industrial Organization, Efficiency, Aggregation, Productivity

Suggested Running Head: Industrial Organization Measure

*Department of Econometrics and Operations Research, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Phone: +31-13-4662365; fax: +31-13-4663280; e-mail: tenRaa@UvT.nl. I am grateful to Thanh Le Phuoc for good research assistance, William Weber for making his data set available, Shawna Grosskopf for literature references, and Knox Lovell for telling me not to ignore entry and exit.

1 Introduction

If relatively productive firms grow relatively fast, an industry will improve its performance, even when no firm exhibits technical change or efficiency change. The industrial organization changes for the better and contributes to performance. The effect is positive for industries where winners are picked and negative for industries where losers are protected. In this paper the performance of an industry is assessed in terms of technology, efficiency, and its industrial organization. The theory is developed in a nonparametric setting, which merely requires input and output information. Nonetheless I establish the link with the Solow residual, which normally requires a production function. In discrete time I set up the Malmquist variants of technical and efficiency change and capture the industrial organization effect.

In neoclassical economics, particularly Solovian growth theory (1957), it is customary to assume perfect competition and no externalities. The economic equilibrium is efficient and any improvement in output/input ratios can be ascribed to technological progress. The assumption of perfect competition is released in more micro-economic approaches to productivity, ever since Debreu's (1951). His coefficient of resource utilization measures the inefficiency in a static economy and has been commingled with Solow's residual in productivity analyses, particularly Data Envelopment Analysis and Stochastic Frontier Analysis. The insights of Debreu and Solow have been married by showing that productivity growth encompasses not only technical change, but also efficiency change. Roughly speaking, an economy may improve its performance by shifting out its frontier or by approaching it. While this decomposition is crystal clear at the micro level and has been applied to the macro level (Färe et al., 1994), Blackorby and Russell (1999) have shown that things do not add up except under restrictive conditions. Indeeed, aggregating industry productivities to total productivity growth, Jorgenson, Ho, and Stiroh (2003) capture allocative efficiency changes in their formula (53). In the present paper I show that the industrial organization effect reflects the bias in the aggregation literature (ten Raa, 2005) and provide a framework for the measurement of all performance components: technical change, efficiency change, and industrial organization.

The next section introduces the concept of benchmarking a firm against the best practices in an industry, by means of a linear program. The same concept is applied to benchmark the industry against the best practices. Section 3 uses the aggregation bias to measure the inefficiency of an industrial organization. Since efficiency is a function of both the object to be benchmarked and the reference benchmark, time changes may be traced through either argument of the function. Section 4 shows that the change transmitted through the object itself is productivity growth and that the change transmitted through the reference industry is technical change. Productivity growth is thus shown to be the sum of efficiency change and technical change and discrete time approximations are presented in section 5, including the industrial organization effect. Section 6 applies the theory to measure the evolution of the industrial organization of Japanese banking. Section 7 shows how to accommodate the entry and exit of firms in industrial organization measurement. Section 8 concludes.

2 Firms and industry efficiency

Denote firm i's input vector by x^i and its output vector by y^i, $i = 1, ..., I$. Input and output vectors may have different dimensions. For example, inputs can be labor, capital, and land, while outputs may be numerous goods and services. Some commodities can be both input and output. The *industrial organization* is identified with the allocation $(x^i, y^i)_{i=1,...,I}$, which is denoted briefly by (x, y). If $I = 1$, the industrial organization is a monopoly; if $I = 2$, it is a duopoly. If y is a diagonal matrix, we have monopolistic competition. If x is a row vector, we have an input price taking industry, for which inputs can be aggregated to 'cost.' The efficiency of a firm is determined by *benchmarking* firm's structure (x^i, y^i) against the industrial organization (x, y). This is a comparison between the actual output level and the best practice output level achievable with the available input vector. The idea is to reallocate the input, x^i, over all the activities $j = 1, ..., I$, with intensities θ_j, as to inflate the output, y^i, by an expansion factor $1/\varepsilon$:

$$\max_{\varepsilon, \theta_j \geq 0} 1/\varepsilon : \sum \theta_j x^j \leq x^i, y^i/\varepsilon \leq \sum \theta_j y^j \tag{1}$$

Here it is assumed that activity (x^i, y^i) can be run with constant returns to scale.[1] Let ε^i solve primal program (1).[2] The expanded y^i/ε^i is the *potential* output of firm i, using the best practice technologies. If $\varepsilon^i = 0.9$, firm i could produce a factor $1/0.9 = 1.11$ or 11% more. If $\varepsilon^i = 1$, potential output is no more than actual output and firm i is said to be fully efficient. In general, ε^i is a number between 0 and 1 which indicates the *firm efficiency* for firm i. The best practice firms or benchmarks relevant to firm i are signalled by $\theta_j > 0$ in program (1).

Denote the shadow prices of the constraints in (1) by w^i and p^i, for the inputs and outputs, respectively. They solve the dual program:[3]

$$\min_{p, w \geq 0} wx^i : py^j \leq wx^j, py^i = 1, \text{ all } j \tag{2}$$

By the main theorem of linear programming the primal and dual programs have equal solution values:

$$1/\varepsilon^i = w^i x^i \tag{3}$$

Substituting the price normalization constraint of program (2) in equation (3), the efficiency of firm i becomes:

[1] Alternatively, one might contract the input. However, under constant returns to scale output and input based benchmarking are equivalent.

[2] Notice the program is linear in the nonnegative variables $1/\varepsilon, \theta_j$.

[3] The price normalization constraint features no slack, because the non-negativity constraint for $1/\varepsilon$ is non-binding (as $1/\varepsilon = 1$ is feasible by choice of $\theta_i = 1$ and $\theta_j = 0, j \neq i$).

$$\varepsilon^i = p^i y^i / w^i x^i \tag{4}$$

The efficiency is the ratio of the value of output to the value of input at internal, firm accounting prices. The prices are firm specific for two reasons: (i) potential output of firm i has idiosyncratic commodity proportions; (ii) there are multiple inputs. The first cause is straightforward. If the output mix of a firm is relatively intensive in terms of some input, the shadow price of that input will be high. It follows that firm efficiency is a *private* measure. The second cause is deep. If there is essentially one input, as for an industry that is input price taking, then the shadow prices of the outputs can be shown to be independent of the firm's output mix (the Samuelson substitution theorem, ten Raa, 1995), ensuring perfect agreement between private and social values.

Shadow input prices are high.[4] By the dual constraints in program (2), no firm makes positive profit at shadow prices. Benchmarks break even (by the phenomenon of complementary slackness, see ten Raa, 2006) and inefficient firms incur a loss. This observation confirms that efficiency measure (4) is a number between 0 and 1.

We now apply the apparatus to the efficiency of the industry. Johansen (1972) defined potential industry output as a function of total input. Following Färe and Grosskopf's (2004) extension, the idea is to reallocate the inputs of all firms, industry input $\overline{x} = \sum x^i$, as to inflate the aggregate, industry output, $\overline{y} = \sum y_i$, by an expansion factor $1/\varepsilon$. Program (1) thus becomes:

$$\max_{\varepsilon, \theta_j \geq 0} 1/\varepsilon : \sum \theta_j x^j \leq \overline{x}, \overline{y}/\varepsilon \leq \sum \theta_j y^j \tag{5}$$

Let $\overline{\varepsilon}$ solve program (5). It is a number between 0 and 1 which indicates the *industry efficiency*. The best practice firms or *benchmarks* relevant to the industry are signalled by $\theta_j > 0$ in program (5). Denote the shadow prices of the constraints in program (5) by $\overline{w}$ and $\overline{p}$, for the inputs and outputs, respectively. They solve the dual program:

$$\min_{p,w \geq 0} w\overline{x} : py^j \leq wx^j, p\overline{y} = 1, \text{ all } j \tag{6}$$

Analogous to equation (3), potential output increases by the following factor:

$$1/\overline{\varepsilon} = \overline{wx} \tag{7}$$

Analogous to equation (4), industry efficiency becomes:

$$\overline{\varepsilon} = \overline{py}/\overline{wx} \tag{8}$$

The efficiency is the ratio of the value of output to the value of input. Notice that the price normalization is a wash, since prices are in the numerator and the denominator of formula (8).

[4] This is paradoxal, since the dual program minimizes the value of the inputs. The constraint, however, lifts the value of the inputs over the value of the outputs, in fact to the level of potential output. And the latter is maximized by the primal program.

3 An industrial organization measure

Proposition 1 establishes a relationship between the efficiency of the industry and the efficiencies of the firms.

Proposition 1. Industry efficiency is *less* than the market share weighted harmonic mean of the firm efficiencies: $\bar{\varepsilon} \leq 1/\sum \frac{s^i}{\varepsilon^i}$, where $s^i = \bar{p}y^i/\overline{py}$ are the market shares evaluated at the prices determined by dual program (6).

Proof. In the dual program (2), consider the socially optimal prices $\bar{p}/\bar{p}y^i$ and $\bar{w}/\bar{p}y^i$ (which need *not* be privately optimal). The denominator has been chosen as to fulfil the price normalization constraint in program (2) and the inequality constraint carries over from program (6). In short, these prices are feasible with respect to program (2). But by their suboptimality (in this private minimization program), $(\bar{w}/\bar{p}y^i)x^i \geq w^i x^i$ or $\bar{w}x^i \geq \bar{p}y^i w^i x^i = \bar{p}y^i/\varepsilon^i$, using equation (3). Summing and invoking equation (7) and the price normalization constraint of (6), we obtain $1/\bar{\varepsilon} = \bar{w}x^I = \bar{w}\sum x^i \geq \sum \bar{p}y^i/\varepsilon^i = \sum s^i/\varepsilon^i$. Inverting, industry efficiency becomes $\bar{\varepsilon} \leq 1/\sum \frac{s^i}{\varepsilon^i}$. Q.E.D.

Corollary 1. If the relative private and social prices are equal, then the industry efficiency equals the market share weighted harmonic mean of the firm efficiencies.

Proof. Let the relative private and social prices be equal: $w^i = \lambda_i \bar{w}$ and $p^i = \lambda_i \bar{p}$. Then in the proof of Proposition 1, $\bar{p}/\bar{p}y^i = p^i/p^i y^i = p^i$ and $\bar{w}/\bar{p}y^i = w^i/p^i y^i = w^i$, by the normalization constraint in program (2). Hence they are privately optimal and the inequalities in the proof of Proposition 1 are binding. Consequently, $\bar{\varepsilon} = 1/\sum \frac{s^i}{\varepsilon^i}$. Q.E.D.

The reason that industry efficiency is less than mean firm efficiency is that the industrial organization is suboptimal. It is a form of allocative inefficiency. Firms better be split or merged, specialize or diversify. The optimal industrial organization is determined by the benchmarks in program (5). Suboptimality is signalled by a distortion between private and social prices (Corollary 1). The efficiency of the industrial organization can thus be measured by the ratio of the industry efficiency to the mean firm efficiency, or, using Proposition 1:

Definition 1. The *efficiency of an industrial organization*, (x, y), equals $\varepsilon^{IO} = \bar{\varepsilon}\sum s^i/\varepsilon^i$, where s^i are the market shares evaluated at the prices determined by dual program (6).

Notice that by Proposition 1 the efficiency of an industrial organization is indeed a number between 0 and 1, with the latter value representing full efficiency according to Corollary 1.

Examples. **1.** Consider an industry with equally efficient firms: $\varepsilon^i = \varepsilon$. Then by Proposition 1, $\bar{\varepsilon} \leq 1/\sum \frac{s^i}{\varepsilon^i} = \varepsilon$. Hence industry efficiency is less than firm efficiency. The efficiency of the industrial organization is $\varepsilon^{IO} = \bar{\varepsilon}/\varepsilon$.

2. Consider an industry that produces a single good from labor and capital. Three firms each produce one unit of output. Firm 1 uses just one unit of labor, firm 2 uses just one unit of capital, and firm 3 uses 1/3 units of both commodities. Since firm 1 has labor only, the technologies of firms 2 and 3 (which employ capital) are of no use. There is no potential increase of its output. The same conclusion holds for firm 2. Firm 3 could reallocate its labor and capital to the technologies employed by firms 1 and 2, respectively, but its output would go down from 1 to 2/3. Hence no firm has scope for an increase in output. All potential outputs are equal to the observed outputs, all firms are 100% efficient. The industry, however, is not efficient. If firms 1 and 2 would merge and adopt the technology of firm 3, the new firm would be three times as big as firm 3, hence produce three units of output, which is one more than they produce using their own technologies. Potential output is four units (instead of three), so that the expansion factor is 4/3 and, therefore, the industry efficiency is 3/4 or only 75%. The efficiency of the industrial organization is $75/100 = 0.75$ or 75%. The industry would do better if the two specialized firms would merge.

3. It is straightforward to construct an example where the industry would do better if a firm were broken up: Simply substitute diseconomies of scope for the economies of scope in Example 2, by letting firm 3 use 2/3 units of both inputs.

4. Add a fourth firm to Example 2 which has the same inputs as firm 3, but only 1/2 a unit of output. Clearly, firm 4 could produce a full unit of output (adopting the technology of firm 3). Its efficiency is 50%. In the present example, the outputs are 1, 1, 1, 0.5. The market shares are 2/7, 2/7, 2/7, 1/7. The firm efficiencies are 100%, 100%, 100%, 50%. The harmonic mean is $1/\sum s^i/\varepsilon^i = 1/(\frac{2/7}{1.00} + \frac{2/7}{1.00} + \frac{2/7}{1.00} + \frac{1/7}{0.50})$ or 87.5%. For the industry potential output is three for firms 1 and 2 jointly (see Example 2) and one for firms 3 and 4 each, hence five in total (instead of three and a half), so that the expansion factor is 5/3.5 and, therefore, the industry efficiency is 3.5/5 or only 70%. The efficiency of the industrial organization is $70/87.5 = 0.8$ or 80%.

The upshot for performance analysis is the following.

Corollary 2. Industry efficiency is the product of (market share weighted harmonic) mean firm efficiency and the efficiency of the industrial organization.

Corollary 2 will enable us to refine the decomposition of productivity growth in technical change and efficiency change.

4 Productivity growth

In the previous section I interrelated the efficiency levels of firms with that of the industry in a snapshot. Now time is introduced by subscripting inputs and outputs, as well as the derived constructs, using the symbol t. Firm i has input and output vectors x_t^i and y_t^i, $i = 1, ..., I$. By benchmarking, y_t^i/ε_t^i is derived, the potential output of firm i. Its efficiency is indicated by ε_t^i, a number between 0 and 1. As a percentage, efficiency change is:

$$EC_t^i = \frac{d}{dt}\varepsilon^i{}_t/\varepsilon_t^i \tag{9}$$

It is important to understand that efficiency may change for reasons of internal, firm organization *and* for reasons of external, industrial organization. If the production possibilities of the industry remain constant, but firm i improves its output/input ratio (or productivity), thus getting closer to the production possibility frontier, the better internal organization yields positive efficiency change. In this case, its productivity growth equals efficiency change, while technical change is zero. If firm i has a constant output/input ratio, but the production possibility frontier of the industry shifts out, the better external organization implies negative efficiency change for the firm. In this case efficiency change and technical change cancel out and the firm has zero productivity growth. In either case, efficiency change and technical change sum to the firm's productivity growth:

$$PG_t^i = EC_t^i + TC_t^i \tag{10}$$

I have defined efficiency change, but not yet technical change. Technical change manifests itself as a shift of the production possibility frontier. At each point of time, the frontier is determined by the industrial organization (x_t, y_t). The efficiency of firm i is determined by program (1). Its input-output pair, (x_t^i, y_t^i), is benchmarked against (x_t, y_t). Formally, program (1) determines the efficiency of firm i as a function of (x_t^i, y_t^i) and (x_t, y_t). Hence we may write $\varepsilon_t^i = e((x_t^i, y_t^i), (x_t, y_t))$, where mapping e summarizes the efficiency program. Notice that the program that determines the efficiency of the industry, (5), has *precisely* the same structure as that for the firms, hence the *same* mapping e governs the relationship between the data and industry efficiency. The only difference is that it benchmarks the industry input-output pair, $(\overline{x}_t, \overline{y}_t)$. Consequently, program (5) may be written as $\overline{\varepsilon}_t = e((\overline{x}_t, \overline{y}_t), (x_t, y_t))$, with the same mapping e. Mapping e has two arguments, the input-output pair that is benchmarked, (x_t^i, y_t^i) in case of the firm, and the industry constellation that determines the frontier, (x_t, y_t). Notice that the structure of program (1) or (5) is independent of time, so that time does not enter the function as a separate argument. Denote the two partial derivatives of the mapping by e_1 and e_2.[5] By total differentiation, the efficiency change of firm i is:

[5] Both are row vectors, because either argument has a number of components: the number of commodities and the I-fold of the number of commodities, respectively. Moreover, because

$$\begin{array}{l}\frac{de}{dt}((x_t^i,y_t^i),(x_t,y_t))/e((x_t^i,y_t^i),(x_t,y_t)) \\ = e_1((x_t^i,y_t^i),(x_t,y_t))\frac{d}{dt}(x_t^i,y_t^i)/e((x_t^i,y_t^i),(x_t,y_t)) \\ + e_2((x_t^i,y_t^i),(x_t,y_t))\frac{d}{dt}(x_t,y_t)/e((x_t^i,y_t^i),(x_t,y_t))\end{array} \tag{11}$$

The measurement of technical change is subtle. If firm i stays put–(x_t^i, y_t^i) = constant–but potential output increases, there must be technical progress. Now an increase in potential output, $1/\varepsilon_t^i$, is equivalent to a decrease in efficiency, ε_t^i. Hence a *negative* second partial derivative (which captures the external effect) indicates technical progress and, therefore, *technical change* is measured by:

$$TC_t^i = -e_2((x_t^i,y_t^i),(x_t,y_t))\frac{d}{dt}(x_t,y_t)/e((x_t^i,y_t^i),(x_t,y_t)) \tag{12}$$

Finally, *productivity growth* of firm i ought to be defined irrespective the shift of the production possibility frontier; it is the own effect of a firm on its efficiency:

$$PG_t^i = e_1((x_t^i,y_t^i),(x_t,y_t))\frac{d}{dt}(x_t^i,y_t^i)/e((x_t^i,y_t^i),(x_t,y_t)) \tag{13}$$

Here input and output changes are prized by the marginal products of the firm. In other words, PG_t^i is the Solow residual of firm i:

Proposition 2. Productivity growth is measured by the Solow residual: $PG_t^i = (p_t^i \frac{d}{dt} y_t^i - w_t^i \frac{d}{dt} x_t^i)/w_t^i x_t^i$

Proof. The proof is by duality analysis. Mapping e's first argument, (x_t^i, y_t^i), lists the bounds in program (1). Now the partial derivatives of the objective value, $1/\varepsilon_t^i$, with respect to the bounds are the shadow prices, w_t^i and $-p_t^i$. The different signs reflect the opposite places of the terms in the inequalities in the program. Consequently, $\frac{-1}{\varepsilon_t^{i2}}\frac{\partial \varepsilon_t^i}{\partial x_t^i} = w_t^i$ and $\frac{-1}{\varepsilon_t^{i2}}\frac{\partial \varepsilon_t^i}{\partial y_t^i} = -p_t^i$. Substituting these in expression (13), we obtain $PG_t^i = (-w_t^i, p_t^i)\varepsilon_t^{i2}\frac{d}{dt}(x_t^i, y_t^i)/\varepsilon_t^i = (p_t^i \frac{d}{dt} y_t^i - w_t^i \frac{d}{dt} x_t^i)/w_t^i x_t^i$, by equation (3). Q.E.D.

Summarizing, efficiency change is defined by (9), technical change by (12), productivity growth by (13), and the former two sum to the latter by equation (11), which confirms our intuitive equation (10).

Things look only slightly different at the level of the industry. Now industry input and output, (x_t^I, y_t^I), are benchmarked against the frontier. The productivity growth of the industry is:

$$\overline{PG}_t = e_1((\overline{x}_t,\overline{y}_t),(x_t,y_t))\frac{d}{dt}(\overline{x}_t,\overline{y}_t)/e((\overline{x}_t,\overline{y}_t),(x_t,y_t)) \tag{14}$$

of the possibility of jumps, the derivatives must be generalized to signed measures. The partial derivatives capture the shadow prices, as revealed by the ensueing analysis.

This expression is basically a summation of the firm productivity growth rates, (13), with the modification that private shadow prices have been replaced by social values. This difference constitutes precisely the aggregation bias uncovered by ten Raa (2005). The same difference between private and social valuations causes a bias in the aggregation of technical change, (12), but here it is a minor phenomenon, specific to the nonparametric approach. Consequently, the productivity aggregation bias is basically equal to the efficiency aggregation bias, or, invoking Corollary 2, the industrial organization effect. The next section will explicate the role of industrial organization in the performance measure of productivity.

5 Malmquist indices

In discrete time a fascinating thought construct is to benchmark a firm against the industry at another period. One may hope that the efficiency of a firm benchmarked against the industry in the next period, $e((x_t^i, y_t^i), (x_{t+1}, y_{t+1}))$, is low. The basic idea of the Malmquist productivity index is to trace firm i from period t to $t+1$ and to measure the change in efficiency relative to a fixed benchmark. For example, benchmarking against the second period yields
$e((x_{t+1}^i, y_{t+1}^i), (x_{t+1}, y_{t+1})) - e((x_t^i, y_t^i), (x_{t+1}, y_{t+1}))$. This difference expression is a discrete time version of the numerator of productivity growth expression (13). Efficiency change contributes to the first term and technical change to the second term. The discrete time frame prompts two minor modifications. First, Malmquist indices are ratios instead of level changes, so that the difference expression turns $\frac{e((x_{t+1}^i, y_{t+1}^i), (x_{t+1}, y_{t+1}))}{e((x_t^i, y_t^i), (x_{t+1}, y_{t+1}))}$.[6] Second, since one could just as well benchmark against the previous period, which would yield $\frac{e((x_{t+1}^i, y_{t+1}^i), (x_t, y_t))}{e((x_t^i, y_t^i), (x_t, y_t))}$, the geometric average of the two possibilities is taken. In short, the Malmquist productivity index is defined by (Färe et al. 1989):

$$M_t^i = \sqrt{\frac{e((x_{t+1}^i, y_{t+1}^i), (x_t, y_t))}{e((x_t^i, y_t^i), (x_t, y_t))} \cdot \frac{e((x_{t+1}^i, y_{t+1}^i), (x_{t+1}, y_{t+1}))}{e((x_t^i, y_t^i), (x_{t+1}, y_{t+1}))}} \tag{15}$$

Incidentally, it is straightforward to recover the decomposition in efficiency change and technical change (ten Raa and Shestalova, 2006). Simply rewrite index (15) as follows:

$$M_t^i = \frac{e((x_{t+1}^i, y_{t+1}^i), (x_{t+1}, y_{t+1}))}{e((x_t^i, y_t^i), (x_t, y_t))} \cdot \sqrt{\frac{e((x_t^i, y_t^i), (x_t, y_t))}{e((x_t^i, y_t^i), (x_{t+1}, y_{t+1}))} \cdot \frac{e((x_{t+1}^i, y_{t+1}^i), (x_t, y_t))}{e((x_{t+1}^i, y_{t+1}^i), (x_{t+1}, y_{t+1}))}} \tag{16}$$

The first quotient in decomposition (16) measures the increase in efficiency from time t to time $t+1$. The remainder, the square root, contains two quotients

[6] At a first order Taylor approximation this is equivalent to a change of variable, to the exponential function. For example, a growth rate of 1% will yield an index of 1.01. It is customary though to stick to the percentage format in reporting Malmquist indices.

in which the firm is fixed (at time t, respectively $t+1$), but the benchmark shifts; this measures technical change.

Turning from firm i to the industry, benchmark industry input and output against the frontier. Comparison with the firm index (15) shows that the industry Malmquist productivity index becomes:

$$\overline{M}_t = \sqrt{\frac{e((\overline{x}_{t+1}, \overline{y}_{t+1}), (x_t, y_t))}{e((\overline{x}_t, \overline{y}_t), (x_t, y_t))} \cdot \frac{e((\overline{x}_{t+1}, \overline{y}_{t+1}), (x_{t+1}, y_{t+1}))}{e((\overline{x}_t, \overline{y}_t), (x_{t+1}, y_{t+1}))}} \tag{17}$$

Proposition 3. The Malmquist productivity index aggregates the change in the efficiency of the industrial organization, firm efficiency changes, and technical change:

$$\overline{M}_t = \frac{\varepsilon_{t+1}^{IO}}{\varepsilon_t^{IO}} \cdot \frac{\sum s_t^i / e((x_t^i, y_t^i), (x_t, y_t))}{\sum s_{t+1}^i / e((x_{t+1}^i, y_{t+1}^i), (x_{t+1}, y_{t+1}))} \cdot$$
$$\cdot \sqrt{\frac{e((\overline{x}_t, \overline{y}_t), (x_t, y_t))}{e((\overline{x}_t, \overline{y}_t), (x_{t+1}, y_{t+1}))} \cdot \frac{e((\overline{x}_{t+1}, \overline{y}_{t+1}), (x_t, y_t))}{e((\overline{x}_{t+1}, \overline{y}_{t+1}), (x_{t+1}, y_{t+1}))}}$$

The first quotient measures the change in the efficiency of the industrial organization. Firm efficiencies are aggregated in the second quotient by the market share weighted harmonic mean and market shares are evaluated at the shadow prices of the industry efficiency program (5). The square root measures technical change.

Proof. Apply formula (16) to the industry and substitute, using Definition 1, for $e((\overline{x}_t, \overline{y}_t), (x_t, y_t)) = \varepsilon_t^{IO} / \sum s_t^i / e((x_t^i, y_t^i), (x_t, y_t))$ and similar for $e((\overline{x}_{t+1}, \overline{y}_{t+1}), (x_t, y_t))$. Q.E.D.

6 Application

Consider the Japanese banks ($i = 1, ..., I = 136$) over a five year period ($t = 1992, ..., 1996$).[7] There are three inputs (labor, capital, and funds from customers) and two outputs (loans and other investments). Formally we have a panel of inputs and outputs, (x_t^i, y_t^i). For the four transitions between periods Thanh Le Phuoc has computed the dynamic performance measure of productivity growth, and, applying Proposition 3, its decomposition in the industrial organization effect, firms efficiency change and technical change. The results are in Table 1.[8]

[7] Fukuyama and Weber (2002) kindly made available their data. The data were obtained by extracting Nikkei's data tape of bank financial statements. Six banks had missing data and were excluded. These were Akita Akebono, Bank of Tokyo, Hanwa, Hyogo, Midori, and Taiheiyo.

[8] When Malmquist indices and its components are properly reported as fractions of the order 1, the product of the components equals total productivity growth. When reported as percentages, the components sum to total factor productivity growth, up to a first order Taylor approximation. This and rounding errors explain why not all row figures add.

Table 1.
The three contributions to the performance of the Japanese banking industry

Period	Industrial Organization	Firms Efficiency	Technical Change	Total Productivity
1992-1993	0.07%	-0.51%	0.52%	0.08%
1993-1994	0.57%	0.31%	-0.58%	0.29%
1994-1995	-0.45%	0.35%	1.28%	1.18%
1995-1996	0.23%	0.70%	2.19%	3.15%
Total, annualized	0.11%	0.21%	0.85%	1.17%

The results permit a diagnosis of the Japanese banking industry. In the mid 1990s Japanese banking showed a solid performance of 1.17% productivity growth per year, much due to a final sprint. The bulk, in fact a share of 73%, was due to technical change, such as advances in electronic banking. The second biggest chunk, in fact a share of 18%, was due to efficiency change at the bank level, such as the spread of ATMs. Last and least, industrial reorganization accounts for 9% of the Japanese banking productivity growth. Various explanations can be advanced to understand these different contributions, such as R&D, competitive pressure, and changes in bankruptcy procedures. Many observers feel that there is scope for a bigger role of the industrial reorganization of Japanese banking. True or not, the first task seems to be the measurement of its share in productivity growth. At least that can now be ticked off the research agenda.

7 Entry and exit

An important source of productivity growth is the entry and exit of relatively productive and relatively unproductive firms.[9] An analysis of this phenomenon requires the consideration of different numbers of firms at the beginning and end of a period. Thus, denote the number of firms at time t by I_t. Moreover, we must consider entrants and exitors. There is a slight time asymmetry. The newness of entrants' technologies means they were not available in the past, but the oldness of exitors' technologies does not mean they are no longer available in the future. Old technologies become obsolete economically (unprofitable). This asymmetry is modelled by means of sequential Malmquist indices; see Tulkens and Vanden Eeckaut (1995). While entrants are modelled as new firms, exitors continue to exist, but become dormant–with zero output and input levels. Denote the number of entrants by E_t. Then

[9] This is according to Knox Lovell.

$I_{t+1} = I_t + E_t$ and the partition of firms at time $t+1$ in incumbents and entrants reads $\mathcal{I}_{t+1} = \mathcal{I}_t \cup \mathcal{E}_t = \{1, ..., I_t, I_t + 1, ..., I_t + E_t\}$. Denote the total input-output combinations of incumbents by $(\overline{x}_{t+1}, \overline{y}_{t+1})^{\mathcal{I}} = (\overline{x}^{\mathcal{I}}_{t+1}, \overline{y}^{\mathcal{I}}_{t+1})$ and similar for the entrants. Similarly, the industrial organization at time $t+1$ can be written $(x_{t+1}, y_{t+1}) = ((x_{t+1}, y_{t+1})^{\mathcal{I}}, (x_{t+1}, y_{t+1})^{\mathcal{E}})$. Industry efficiency becomes $\overline{\varepsilon}_{t+1} = e((\overline{x}_{t+1}, \overline{y}_{t+1})^{\mathcal{I}} + (\overline{x}_{t+1}, \overline{y}_{t+1})^{\mathcal{E}}; (x_{t+1}, y_{t+1})^{\mathcal{I}}, (x_{t+1}, y_{t+1})^{\mathcal{E}})$, where the first argument is equal to $(\overline{x}_{t+1}, \overline{y}_{t+1})$.

By Proposition 1, industry efficiency is less than the market share weighted harmonic mean of the firm efficiencies. It is straightforward to extend this aggregation result to groups of firms. In other words, industry efficiency is less than the market share weighted harmonic mean of the group of incumbent firms and the group of entrants. The difference is the inefficiency involved with the suboptimal balance between entrants and incumbents. Incumbents' efficiency is $\overline{\varepsilon}^{\mathcal{I}}_{t+1} = e((\overline{x}_{t+1}, \overline{y}_{t+1})^{\mathcal{I}}; (x_{t+1}, y_{t+1})^{\mathcal{I}}, (x_{t+1}, y_{t+1})^{\mathcal{E}})$ and entrants' efficiency is similar. Denote the market share for incumbents by $s^{\mathcal{I}}_{t+1} = \overline{p}_{t+1} y^{\mathcal{I}}_{t+1} / \overline{p}_{t+1} \overline{y}_{t+1}$, where the prices are determined by the year $t+1$ version of dual program (6), and similar for the entrants. Then $\overline{\varepsilon}_{t+1} \leq 1/(\frac{s^{\mathcal{I}}_{t+1}}{\overline{\varepsilon}^{\mathcal{I}}_{t+1}} + \frac{s^{\mathcal{E}}_{t+1}}{\overline{\varepsilon}^{\mathcal{E}}_{t+1}})$.

Definition 2. The *efficiency of entry*, $((x,y)^{\mathcal{I}}, (x,y)^{\mathcal{E}})$, equals $\varepsilon^E = \overline{\varepsilon}(\frac{s^{\mathcal{I}}}{\overline{\varepsilon}^{\mathcal{I}}} + \frac{s^{\mathcal{E}}}{\overline{\varepsilon}^{\mathcal{E}}})$, where $s^{\mathcal{I}}$ and $s^{\mathcal{E}}$ are the market shares of incumbents and entrants evaluated at the prices determined by dual program (6).

Proposition 4. The Malmquist productivity index aggregates the efficiency of entry, incumbent and entrant efficiency changes, and technical change:

$$\overline{M}_t = \varepsilon^E_{t+1} \cdot \frac{\frac{1}{\overline{\varepsilon}_t}}{\frac{s^{\mathcal{I}}_{t+1}}{\overline{\varepsilon}^{\mathcal{I}}_{t+1}} + \frac{s^{\mathcal{E}}_{t+1}}{\overline{\varepsilon}^{\mathcal{E}}_{t+1}}} \cdot \sqrt{\frac{e((\overline{x}_t, \overline{y}_t), (x_t, y_t))}{e((\overline{x}_t, \overline{y}_t), (x_{t+1}, y_{t+1}))} \cdot \frac{e((\overline{x}_{t+1}, \overline{y}_{t+1}), (x_t, y_t))}{e((\overline{x}_{t+1}, \overline{y}_{t+1}), (x_{t+1}, y_{t+1}))}}$$

The first factor measures the dynamic industrial organization effect. Incumbent and entrant efficiency changes are aggregated in the second quotient by the market share weighted harmonic mean and market shares are evaluated at the shadow prices of the industry efficiency program (5). The square root measures technical change.

The middle factor accounts for the efficiency change of incumbent and entrants at an aggregated level, including not only firm efficiency changes but also the (static) industrial organization effect. This detail can be inserted as follows. Application of Corollary 2, $1/\overline{\varepsilon} = \frac{\sum s^i/\varepsilon^i}{\varepsilon^{IO}}$, to the incumbents and entrants transforms the middle factor in Proposition 4 to $\frac{\frac{1}{\overline{\varepsilon}_t}}{\frac{s^{\mathcal{I}}_{t+1}}{\overline{\varepsilon}^{\mathcal{I}}_{t+1}} + \frac{s^{\mathcal{E}}_{t+1}}{\overline{\varepsilon}^{\mathcal{E}}_{t+1}}} = \frac{\sum s^i_t/\varepsilon^i_t}{\varepsilon^{IO}_t} / \left(s^{\mathcal{I}}_{t+1} \frac{\sum_{\mathcal{I}} s^i_{t+1}/\varepsilon^i_{t+1}}{\varepsilon^{IO\mathcal{I}}_{t+1}} + s^{\mathcal{E}}_{t+1} \frac{\sum_{\mathcal{E}} s^i_{t+1}/\varepsilon^i_{t+1}}{\varepsilon^{IO\mathcal{E}}_{t+1}} \right)$. This expression is a combination of firm efficiency changes, $\varepsilon^i_{t+1}/\varepsilon^i_t$, entrants efficiencies, ε^i_{t+1}, and industrial organization effects, $\varepsilon^{IO\mathcal{I}}_{t+1}/\varepsilon^{IO\mathcal{I}}_t$ and $\varepsilon^{IO\mathcal{E}}_{t+1}$.

8 Conclusion

An industry may perform better, in the sense of productivity growth, by technical progress or by efficiency change. Both sources of growth have been decomposed to the firms of an industry, but the aggregation is imperfect. An industry may improve its performance by industrial reorganization as well. The inefficiency of an industrial organization mirrors the bias in the aggregation of the efficiencies of the firms. It may be reduced by reallocations, as come with the picking of winners. The industrial organization effect is measured by the change in the ratio of the industry efficiency to the market share weighted harmonic mean of the firm efficiencies. The dynamic industrial organization effect of entry and exit can be accommodated. All measures, including the efficiency of an industrial organization, can be calculated using only input and output data of the firms.

References

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

[2] Debreu, G.: "The Coefficient of Resource Utilization," *Econometrica* 19, 3, 273-292 (1951)

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

[4] Färe, R., Grosskopf, S., Lindgren, B., Roos, P., "Productivity Development in Swedish Hospitals: a Malmquist Output Index Approach," Discussion paper No. 89-3, Southern Illinois University, Illinois (1989)

[5] Färe, R., Grosskopf, S., Norris, M., Zhang, Z., "Productivity Growth, Technical progress, and Efficiency Change in Industrialized Countries," *American Economic Review* 84, 1, 66-83 (1994)

[6] Fukuyama, H. and W.L. Weber, "Estimating Output Allocative Efficiency and Productivity Change: Application to Japanese Banks," *European Journal of Operational Research*, 137, 177-190, 2002

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

[8] 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)

[9] Solow, R.M. "Technical Change and the Aggregate Production Function," *The Review of Economics and Statistics* 39, 3, 312-320 (1957)

[10] ten Raa, T., "The Substitution Theorem," *Journal of Economic Theory* 66, 2, 632-36 (1995)

[11] ten Raa, T., "Aggregation of Productivity Indices: The Allocative Efficiency Correction," *Journal of Productivity Analysis* 24, 2, 203-09 (2005)

[12] ten Raa, T., *The Economics of Input-Output Analysis*, Cambridge University Press (2006)

[13] ten Raa, T. and V. Shestalova, "Alternative Measures of Total Factor Productivity Growth," Intermediate Input-Output Meetings on Sustainability, Trade & Productivity, Sendai (2006)

[14] Tulkens, H. and P. Vanden Eeckaut, "Non-parametric Efficiency, Progress and Regress Measure for Panel Data: Methodological Aspects," *European Journal of Operational Research* 80, 474-499 (1995)