Chapter 5
Efficiency Measures for Industrial Organization

Thijs ten Raa

5.1 Efficiency and Industrial Organization

If you want to rank firms in an industry, you better use a uniform set of weights for performance components across the firms one compares. This paper explains how potentially conflicting performance aspects can be balanced by assigning rational weights. Here, we distinguish between internal and external (or competitive) benchmarking.

Internal benchmarking helps to spot exemplary business units within big companies, such as hotel chains. For every unit, it spots the relevant benchmarks and suggests cost components that can be cut and potential revenue sources that would boost performance. *External benchmarking* includes the competition in the pool of examples and, therefore, is also called *competitive benchmarking*. It is a more demanding management tool, if only because data are hard to get from competitors, but potentially rewarding. By extending the pool to which you compare, the benchmark will be better and, therefore, the achievement level higher. External benchmarking is relevant to the analysis of industrial organization.

The performance of an industry is more than the sum of the parts. Not only do some firms under perform, but also the reallocation of resources between firms may be suboptimal. Obvious improvements could be achieved by reallocating resources from less to more efficient firms, but more subtle gains can be made by reallocating between efficient firms. Roughly speaking, a better industrial organization would alleviate shortages. In this paper, we measure how much could be gained not only by eliminating firm inefficiencies, but also by a more efficient allocation of resources between the firms. The latter component measures the inefficiency of an industrial organization.

T. ten Raa (✉)
Tilburg School of Economics and Management, Tilburg University,
Tilburg, Netherlands
e-mail: tenRaa@UvT.nl

S.C. Ray et al. (eds.), *Benchmarking for Performance Evaluation*,
DOI 10.1007/978-81-322-2253-8_5

Benchmarking, particularly data envelopment analysis based, is yet to be integrated with other management tools. The airline industry is perhaps the one with the most widespread use of benchmarking. Here, managers routinely measure the performance of their company in comparison with the competition and consumers visit Websites to compare not only prices but also flight statistics, to assess costs and value aspects, such as the reliability of departure and arrival times. Nonetheless, Francis et al. (2005) report that benchmarking is the most popular performance measurement tool in the industry, but that only one airline beefs it up with DEA. The perceived incapacity to value alternative performance components may be a hurdle and we will overcome it.

Most applications of benchmarking are basically alternative performance measurements and rankings, both financial and nonfinancial, and the bulk is based on *output* scores, such as revenues, net earnings, and customer satisfaction. I find it important though to bring in the inputs, if only to separate size effects from true performance scores. I will do so throughout the paper, and we will encounter a number of interesting findings. For example, the use of output to input value *ratios* will emerge as natural performance measures, instead of difference-based level concepts, such as profit (which is revenue minus cost). There is a close connection between the subtle distinction between ratios and difference on one hand and the distinction between the concepts of efficiency, productivity and profitability on the other. All this tends to be a smorgasbord, and it is high time to disentangle the concepts and to clarify which is appropriate for performance measurement and ranking. This paper provides the analysis.

External or competitive benchmarking is applicable both to business units and to corporations. In the former case, comparable intracompany information is required for different companies. In Table 5.1, this case is denoted by the box “Competitive benchmarking.” In the latter case, aggregate company data are compared, which is less demanding. This case is box “External benchmarking” in Table 5.1 and is relevant to industrial organization.

There is an interesting intermediate level of benchmarking, in between internal and external/competitive benchmarking. It is the benchmarking of an industrial organization by measuring it up against its constituent firms. The idea is borrowed from economic theory, which has developed a subtle technique to measure the efficiency of an economy *without* comparing it to other economies. Here, inefficiency encompasses not only suboptimal production of outputs by firms (excessive use of inputs), but also the subtle of form of inefficiency called *allocative* inefficiency. There may be scope for performance improvement by reallocating resources

Table 5.1 A taxonomy of benchmarking

	Reference organization	
Firm	Corporation	Industry
Business unit	Internal benchmarking	Competitive benchmarking
Corporation	Organization benchmarking	External benchmarking

between firms. If so, the firms may be efficient, but the industry is not. This source of inefficiency can be exposed without benchmarking the industry against its foreign competitors, but by benchmarking the industry against its own firms. If we benchmark an industry internally, a subtle conceptual issue emerges and benchmarks are specific to the firm considered. While internal benchmarking and competitive benchmarking are the same ballgame from the point of view of technical analysis, *industrial organization benchmarking* is distinct in that it requires some extra work to remove the dependence.

If we benchmark an industry against its own firms, then we analyze the performance of the industrial organization given the industry's *total* inputs and *total* outputs. We will do so by analyzing how much better the industry could perform if not only each firm operates efficiently, but also the industry's resources are allocated optimally. This problem will be solved by the operations research technique of linear programming, but this time, the benchmark valuations or shadow prices pertain to the industry as a whole. In this way, the results are no longer specific to firms and can be used to *rank* them objectively.

The remainder of this paper is organized as follows. Section 5.2 introduces efficiency measurement at the level of the firm. Section 5.3 discusses the subtle interrelations between efficiency, productivity, and profitability. Section 5.4 discusses ranking and pricing issues. Section 5.5 discusses economies of scale, and Sect. 5.6 presents the ramifications for efficiency measures, including for industries. Section 5.7 decomposes industrial efficiency into firm and organization components and accounts for changes through time, including entry and exit. Section 5.8 concludes.

5.2 Efficiency Measurement of Firms

An *industry* consists of *I firms*. Firm *i* (where $i = 1, \ldots, I$) is a black box that transforms *input* quantities $x_1^i, \ldots, x_k^i$ into *output* quantities $y_1^i, \ldots, y_l^i$. Here, subscript *k* is the number of inputs and subscript *l* is the number of outputs. Superscript *i* indicates the firm of which we take the data. We wish to know at which level a firm operates compared to its full potential (i.e., full efficiency). In other words, we pose the question how much more *could* the firm deliver? To answer this question, we conduct a thought experiment.

We allow the firm to redistribute its inputs over the activities represented by the inputs and outputs of all firms. Hence, imagine firm *i* would employ its inputs to run the activities of firms 1, …, *I* with *intensities* $\theta_1, \ldots, \theta_I$, where *I* is the total number of firms. The firm would need quantities $x_1^1\theta_1 + \cdots + x_1^I\theta_I$ of input 1, …, $x_k^1\theta_1 + \cdots + x_k^I\theta_I$ of input *k*, and produce quantities $y_1^1\theta_1 + \cdots + y_1^I\theta_I$ of output 1, …, $y_l^1\theta_1 + \cdots + y_l^I\theta_I$ of output *l*.

The envisaged operation is feasible if the required inputs do not exceed firm *i*'s available inputs. The operation improves the output level if the hypothetical outputs exceed some multiple of the respective actual outputs of firm *i*. This output multiple is modeled by means of the so-called *expansion* factor, e. The maximum expansion of output of firm i is determined by the following linear program:

$$\begin{array}{lll} \max_{\theta_1,\cdots,\theta_I,\, e \geq 0} e: & & \\ x_1^1\theta_1 + \cdots + x_1^I\theta_I \leq x_1^i, & \cdots, & x_k^1\theta_1 + \cdots + x_k^I\theta_I \leq x_k^i \\ y_1^1\theta_1 + \cdots + y_1^I\theta_I \geq y_1^i e, & \cdots, & y_l^1\theta_1 + \cdots + y_l^I\theta_I \geq y_l^i e \end{array} \tag{5.1}$$

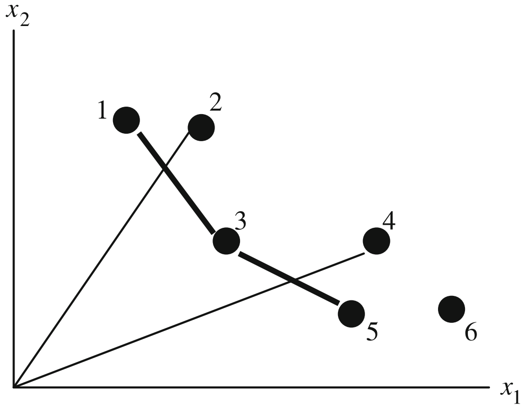
In program (5.1), the expansion factor is maximized subject to the feasibility constraints on the inputs and proportionate expansion of the outputs. *All* that is needed to run benchmarking program (5.1) are the inputs and the outputs of all firms. Implicitly, program (5.1) assumes constant returns to scale. In Sect. 5.5, we will analyze more complicated models.

The assumption of constant returns to scale enables us to normalize the activities. For example, if there are two inputs and one output, it is natural to rescale such that the outputs of all activities are equal to one. After this rescaling, firms differ in their inputs only and we may plot a scatter diagram in input space, see Fig. 5.1.

In Fig. 5.1, firm 4 can reduce its distance to the origin from 100 to 75 %. That way it is moved to the midpoint between firms 3 and 5. Indeed, if we run activities 3 and 5 at intensities 1/2 each, the required inputs are the average of the inputs of firms 3 and 5, which are represented by the midpoint, while the output would be 1/2 + 1/2 = 1, and hence unchanged. This way firm 4 can produce its output using only 75 % of its inputs. The remainder, 25 % of the inputs, can be considered wasted (and be reallocated without reducing output). Because of the constant returns-to-scale assumption, the capacity to produce output with 3/4s of the inputs is equivalent to the capacity to produce 4/3rds of the output with the given input. Application of program (5.1) to unit 4 yields an expansion factor e = 4/3.

Firm 2, also in Fig. 5.1, needs only some 80 % of its inputs if it would employ the techniques of firms 1 and 3; 20 % of its input can be considered waste for this firm. Application of program (5.1) to unit 2 yields an expansion factor e = 5/4. The lower envelop of the observations, the thick lines connecting firms 1, 3, and 5, represents the minimally required inputs for the production of one unit of output. These three units are the *benchmarks*. The distance between the envelope and a firm

Fig. 5.1 An industry with two inputs and a single output. Firms *1*, *3*, and *5* use minimal input. Firms *2*, *4*, and *6* can contract their inputs



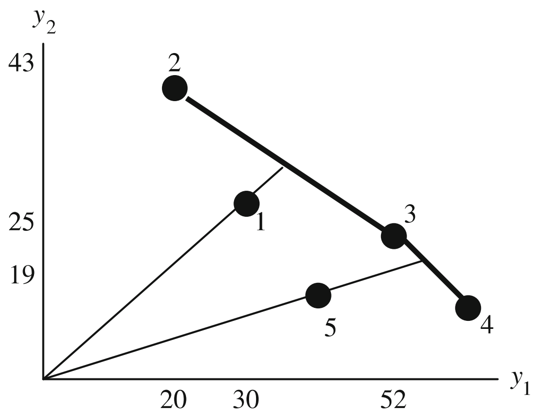


Fig. 5.2 An industry with a single input and two outputs. Firms *2*, *3*, and *4* produce maximal output. Firms *1* and *5* can expand their outputs

measures the unnecessary input, or inefficiency. This technique is called *data envelopment analysis* or DEA for short.

In general, it is impossible to identify an all-purpose best practice or practices. It depends on the unit you benchmark. For example, in Fig. 5.1, there is *no* general purpose best practice. To firm 4, the benchmarks are firms 3 and 5, but to firm 2, the benchmarks are firms 1 and 3. Basically, the relevant best practices must be compatible with the mix of resources of the unit one investigates. In Fig. 5.1, firm 4 is well endowed with the first input. The technique of firm 1, though efficient, requires a lot of the other input and, therefore, is not relevant to firm 4.[1]

Let me move from the example depicted in Fig. 5.1—where two inputs are combined to produce a single output—to the mirror image, where one input, say labor can be used to produce two outputs, such as serving meals and cleaning tables. In this case, it is natural to exploit the assumption of constant returns to scale to the activities such that all inputs are equal to one. Upon this, rescaling firms differ in their outputs only and we may plot a scatter diagram in *output* space, as depicted in Fig. 5.2.

In output space, the most efficient firms are on the North-Eastern frontier. We envelop the data again, but now from above. If firm 1 were to divide its input between the techniques employed by firms 2 and 3, it could produce their average output, which is represented by the midpoint. This way firm 2 can expand its output by some 20 %. The precise output expansion figure is 22 %. Figure 5.2 also shows that firm 5 can produce more by adopting the techniques of firms 3 and 4.

Having developed some intuition what is going on when we calculate wasted inputs, we now proceed with our formal analysis, to pave the way for the determination of the accounting prices, a key tool in the economic analysis. Accounting prices are associated with constraints. Now linear program (1) features $k + l$ ordinary constraints (k for the inputs and l for the outputs) plus I nonnegativity constraints (for the intensities with which the activities are run). Constraints are characterized by the coefficients of the variables. The variables in program (1) are the intensities $\theta_1, \ldots, \theta_I$ *and* expansion factor e. Mind that the input and output

[1] In Fig. 5.1 firm 6 has only one benchmark, namely firm 5.

quantities are *no* variables, but given, fixed data. Formally, they are *coefficients* and, in case of the inputs $(x_1^i, \cdots, x_k^i)$, *bounds*.

Multiplying through the output constraints in (1) by −1, to write them in standard linear programming format (with an ≤–sign), the coefficients in the input constraints are $(x_1^1 \ldots x_1^I\ 0), \ldots, (x_k^1 \ldots x_k^I\ 0)$ and the coefficients in the output constraints are $(-y_1^1 \ldots -y_1^I\ y_1^1), \ldots, (-y_l^1 \ldots -y_l^I\ y_l^1)$, where the last components reflect the right-hand sides of the second line of inequalities in program (5.12). These right-hand side output terms are products of outputs (coefficients) and the expansion variable. They are *not* bounds! The output constraints have no bounds; formally, they are zero. The objective functions coefficients of the variables (θ_1, …, θ_I and e) are $(0 \ldots 0 \quad 1)$, since the intensities θ are tools and only the last variable feeds the objective function. This sums up the formal structure of benchmarking program (1), and we are ready to invoke the accounting prices.

It is customary to denote the accounting prices of the inputs by w_1, …, w_k and of the outputs by p_1, …, p_l. Thus, the dual equation becomes as follows:

$$\begin{array}{lll} 0 & \leq & w_1x_1^1 + \cdots + w_kx_k^1 - p_1y_1^1 - \cdots - p_ly_l^1; \quad \ldots; \\ 0 & \leq & w_1x_1^I + \cdots + w_kx_k^I - p_1y_1^I - \cdots - p_ly_l^I; \\ 1 & \leq & p_1y_1^i + \cdots + p_ly_l^i; \quad w_1, \ldots, w_k \geq 0, p_1, \ldots, p_l \geq 0 \end{array} \tag{5.2}$$

One of the inequalities in (2) is in fact an equality, as I will argue now. The optimal value of expansion factor e is *at least one*, as is seen by the following choice of the intensity variables: $\theta_1 = 1$ and $\theta_2 = \cdots = \theta_I = 0$; this amounts to a simple reproduction of firm i itself. Hence, the nonnegativity constraint for the last variable, e, is *non*binding. The phenomenon of complementary slackness yields that the slack in the dual constraint is zero. Since this is the last component in (2), we obtain the following condition:

$$p_1y_1^i + \cdots + p_ly_l^i = 1 \tag{5.3}$$

Equation (5.3) is the so-called *price normalization* constraint. It resolves arbitrariness in the program that maximizes output (1), as I will explain now. It has to do with the units of measurements. Imagine that all outputs are in kilograms but would be rescaled in metric pounds. Since there are two metric pounds to the kilogram, this would double all y's. The optimal x's would not be affected, nor the relative accounting prices of either the inputs or the outputs. By Eq. (5.3), all the output prices would be halved, precisely as one expects when the unit of measurement is halved. This concludes the explanation of Eq. (5.3).

The other components of dual Eq. (5.2) read as follows:

$$p_1y_1^1 + \cdots + p_ly_l^1 \leq w_1x_1^1 + \cdots + w_kx_k^1; \ldots; p_1y_1^I + p_ly_l^I \leq w_1x_1^I + \cdots + w_kx_k^I \tag{5.4}$$

Dual inequality (4) highlights an important fact: *Accounting prices render all activities unprofitable or zero*. The distinction between negative and zero accounting profits is crucial. It signals which activities are undertaken with positive intensity, in other words which are the benchmarks. It is easy to confirm. By the phenomenon of complementary slackness, we know that if an activity is run with positive intensity, $\theta_i^* > 0$, then there is no slack in the dual constraint, $p_1 y_1^i + p_l y_l^i \leq w_1 x_1^i + \cdots + w_k x_k^i$, and therefore, such activities must break even indeed.

Since accounting prices are shadow prices, we can invoke their marginal productivity interpretation. In the present context, an input price measures by how much the output level could be raised if an additional unit of that input was available and an output price measures how much the overall output *level* could be raised if a unit of that output was gifted. The latter argument is subtle, because in benchmarking, we fix the proportions of the outputs. Hence, if a unit of some output arrives as a gift, the resources must be slightly reallocated away from the production of this output, producing a little more of the other outputs, in order to preserve the proportions. If a price of an output component is higher, than of some other, it means that one unit releases more productive resources. *Output accounting prices measure their values in terms of* resource costs.

In general, accounting prices make some firms *other* than themselves break even. It is extremely interesting to identify them, because they constitute the best available practices. So, focus on the firms for which the associated equations in (4) are binding. It constitutes the set of activities that would be run if the resources available to firm *i* are used optimally. The definition of efficiency is straightforward. Compare a firm to its peers in the industry. More precisely, calculate how much more it could produce by solving program (1). Let the expansion factor be *e*. For example, if *e* = 1.1, it could produce 10 % more and, therefore, it produces only 1/*e* = 0.91 of its potential output. Hence, *efficiency* is simply defined by the inverse value of the expansion factor of benchmarking program (1), 1/*e*. Since the expansion factor is at least one, it follows by sheer arithmetic that *efficiency is a measure between zero and one*. Full efficiency (1/*e* = 1) represents the situation where a firm cannot improve its performance, in other words, it is a leader.

We now use our apparatus to develop a nice alternative expression for efficiency, highlighting the performance of a firm we are interested in. The main tool is the main theorem of linear programming, which equates the value of the objective function with the value of the bounds. In benchmarking, the objective function is the expansion factor, *e*, see program (1). As noted, the value of the bounds in this program is $w_1 x_1^1 + \cdots + w_k x_k^1$. Invoking the price normalization constraint, Eq. (5.3), we conclude that the expansion factor is equal to the accounting cost/revenue ratio:

$$e = (w_1 x_1^i + \cdots + w_k x_k^i)/(p_1 y_1^i + \cdots + p_l y_l^i). \tag{5.5}$$

Formula (5.5) renders the expansion factor robust with respect to price level changes. For example, halving all prices (as in the transition from kilograms to metric pounds) does not affect expression (5.5), because halving the numerator *and* the denominator cancel. Formula (5.5) is important, because it weighs the importance of the inputs and the outputs for efficiency. Obviously, efficiency is increased by reducing inputs or increasing outputs. An input reduction by a unit is more effective if the shadow price is higher and the same holds for an output increase.

Since efficiency has been defined as the inverse of the expansion factor, Eq. (5.5) implies

$$\text{Efficiency} = (p_1 y_1^i + \cdots + p_l y_l^i)/(w_1 x_1^i + \cdots + w_k x_k^i). \tag{5.6}$$

In other words, the performance of a firm is measured by the *revenue/cost ratio* at accounting prices. Equation (5.6) admits a frequently used interpretation of efficiency measurement. Imagine the outputs are produced by a hypothetical, efficient firm. In fact, the efficient inputs are given by the solution of program (5.1), $x_1^1\theta_1 + \cdots + x_1^I\theta_I, \cdots, \quad x_k^1\theta_1 + \cdots + x_k^I\theta_I$, if the outputs are inflated by expansion factor *e*, hence a fraction 1/*e* of these inputs is enough to produce $y_1^1, \ldots, y_l^1$. The efficiency of the hypothetical firm is one. Applying formula (5.6) to the hypothetical unit, we see that it would break even: Revenue equals cost. Here, "cost" is the value of the minimally required inputs to produce the outputs, $y_1^1, \ldots, y_l^1$. Substituting this back in the original formula (5.6), we conclude the following intuitive, yet powerful statement. *Efficiency is the cost/value ratio of inputs*, where "cost" is the value of the *minimally* required inputs, "value" refers to the *actual* inputs, and *both* types of inputs are valued at accounting prices.

5.3 Efficiency, Productivity, and Profitability

We discuss and distinguish some important business concepts which are often confused. An example from the literature is the following quote. Discussing farming enterprises in Finland Yli-Viikari et al. (2002, p. 20) write

> To guarantee the continuation of the production the enterprises have to be profitable. The prerequisite of the profitability is efficiency.

Now particularly in agriculture, there is *no* simple connection between profitability and efficiency. The breakdown of the relationship goes both ways. On the one hand, many inefficient farms are quite profitable, because the prices are maintained at artificially high levels. On the other, efficient firms are not necessarily the most profitable ones. While it is true that a prerequisite for maximum profitability is efficiency, cost cutting is but one way to boost profit. A notorious alternative procedure is the exercise of monopoly power; it deters more efficient firms and creates rents. As Hicks (1935, p. 8) quipped, “The best of all monopoly profits is a quiet life.” In other words, profit need not signal efficiency!

Although there are interrelations between efficiency, productivity, and profitability—and, we will discuss them—the concepts are fundamentally *different*. I will explain the subtleties by means of a simple example. Consider a single-input/single-output industry with two firms, a duopoly. Denote the input quantities by x and the output quantities by y. Use w and p for the prices of the input (the labor wage) and the output (the product price), respectively, and use superscripts to indicate to which firm a symbol pertains: firm 1 or firm 2.

In our discussion, we must distinguish market prices from accounting prices. Market prices are observed and may vary. Some firms negotiate tighter labor conditions than others, and some firms may have shrewder salesmen, extracting higher prices. A well-cited example is someone who “could sell sand to the Arabs.” Indeed, in this situation, *any* price above zero represents market power.

Denote market prices by an underscore and reserve the regular symbols for accounting prices. In short, the symbols for firm 1 are input and output *quantities* x^1 and y^1, input and output *market prices* $\underline{w}^1$ and $\underline{p}^1$, and input and output *accounting prices* w^1 and p^1. The symbols for firm 2 are similar, but with superscript 1 replaced by superscript 2. The techniques implicit in the firm input–output observations (x^1, y^1) and (x^2, y^2) are given by the output/input ratios y^1/x^1 and y^2/x^2. These numbers give the amount of output one can produce with a unit of input and, therefore, constitute the *productivities* of the respective firms. The concept of productivity gets a bit more involved in the presence of multiple inputs and outputs, but this issue can wait. In our simple duopoly, let the first firm be the more productive than the second: $y^1/x^1 > y^2/x^2$. (This is an innocent assumption, because we are free to relabel the firms.) Then, firm 1 can produce no more than it produces, at least under the assumption that the data represent all conceivable practices of production. Firm 2, however, could perform better by adopting the technology of firm 1. That way it would produce y^1/x^1 units per unit of input, and since it commands x^2 inputs, its potential output is $(y^1/x^1)x^2$. By the presumed productivity inequality, this exceeds the actually produced quantity y^2.

The expansion factor, e, measures how much firm 2 could produce relative to what it produces; it equals $e = (y^1/x^1)x^2/y^2$. Now, as we have seen in Sect. 5.2, efficiency is the inverse expansion factor, measuring the actual output as a fraction of potential output: $1/e = y^2/(y^1/x^1)x^2 = (y^2/x^2)/(y^1/x^1)$. Since firm 1 can do no better than using its own technology, its expansion factor equals 1, and therefore, its efficiency is also $1 = (y^1/x^1)/(y^1/x^1)$, to present it in the same format as for firm 2. For both firms, we thus have the following result. *Efficiency equals relative productivity*. Here, productivity is taken relative to the best practice; indeed, firm 1’s productivity features in either denominator.

A further, interesting relationship between the concepts of productivity and efficiency is established in a dynamic setting, where we track the measures through time. If we proceed to the next year, all quantities will be different. Let me assume that the changes are slight, so that firm 1 remains the productivity leader. Imagine that firm 2 has become more efficient. What does it mean? That management does a better job, producing more output per input? Not necessarily. Positive efficiency change only means that relative productivity has increased. One way to boost efficiency is indeed to improve productivity, but another comes with stalling leadership. If the industry leader slips in terms of productivity, the followers get closer, and formally, this shows as positive efficiency change! Think of deteriorating industry conditions. Examples abound. In the mining industry, the quality of the ore lessens as time progresses, simply because the easiest available ores are mined first. Conditions may also deteriorate due to demand, particularly when a product approaches the end of its life cycle, with new substitutes taking over. And conditions are influenced by world markets. For example, if a low-wage country enters the industry and the incumbents are sluggish in making adjustments, such as reallocations to newly developing economies, costs may press harder on smaller numbers of output.

Conversely, without any change in the firm, its efficiency may change. In fact, in a world of technical progress, the best practice productivity increases, and since efficiency is productivity relative to the best practice, it will go down by the denominator effect if there is *no* change in the inputs or outputs of the firm we analyze. The Dutch proverb *standstill is decline* is particularly relevant to the concept of efficiency, because the latter is a *relative* concept. The underlying concept of productivity—output per input—is an absolute concept though. There are many physical examples. The fuel "efficiency" concept of miles per gallon is a productivity measure, indicating how much output one gets per unit of input. If, however, the input and output are of the same dimension, as in energy conversion, we are back in the situation where the "best" is a ratio of one (1 watt of converted energy per watt of basic energy). Here, a ratio equal to one represents the case of no waste, hence full efficiency. Another physical example of a productivity measure is tons harvested (say of rice) per acre. To assess if it signals a good performance, we must know the size relative to the maximum observed magnitude, and *that* is efficiency.

Bring in profitability, a third related concept. In the single-input/single-output duopoly, with firm 1 more productive ($y^1/x^1 > y^2/x^2$), is firm 1 necessarily more *profitable*? Now, we must disentangle a number of issues. First, there is the *size* issue. Firm 1 may enjoy a better margin between revenue and cost per unit of output (because it needs less input) and hence be the more profitable firm per unit of output, but firm 2 may have a bigger market share and hence generate a greater sum of profits. The size effect may outweigh the productivity effect. We correct for it by measuring profit per unit of sales. In other words, instead of profits $\underline{p}y - \underline{w}x$, we compare the profit *rates* $(\underline{p}y - \underline{w}x)/\underline{p}y$.

The second issue is the presence or absence of well functioning markets, both on the input and the output side. If the firms face *common market prices*, both on the input and the output sides, we may compare the profit rate of the more productive firm, which is $(\underline{p}y^1 - \underline{w}x^1)/\underline{p}y^1$, with that of the less productive firm, namely $(\underline{p}y^2 - \underline{w}x^2)/\underline{p}y^2$, and conclude that the former is bigger. The proof of this inequality is easy: The leading terms are equal (namely 1), while the second terms are inverse productivities (x/y) with a common coefficient ($-\underline{w}/\underline{p}$). Inverse productivity is obviously negatively related to productivity, but the minus sign makes the relationship between productivity and profitability per unit sales a positive one.

However—and this takes us to the third issue—the relationship between efficiency and profitability breaks down the moment we drop the single-input/single-output assumption. Figure 5.3 provides a simple example.

In Fig. 5.3, three firms each produce one unit of output using two inputs, namely labor and capital. Firm 2 is inefficient, because it can contract its labor and capital inputs to the midpoint of techniques 1 and 3, as explained in Sect. 5.2. Then, by dividing its inputs between the techniques 1 and 3, it would produce 1/2 + 1/2 = 1 unit of output. However, if capital is very inexpensive, relative to labor, firm 2 will be more profitable than firm 3, simply because it is more economical, using the inexpensive input, capital.

It makes a difference—even for multi-input multi-output industries—if profits are based on accounting prices. It is preferable, because it restores the relationship between efficiency and profit per unit of sales. The key to the analysis is Eq. (5.6), which equates efficiency to the revenue/cost ratio at accounting prices. Simple manipulation modifies that equation into the following:

$$\frac{(p_1 y_1^1 + \cdots + p_l y_l^1) - (w_1 x_1^1 + \cdots + w_k x_k^1)}{p_1 y_1^1 + \cdots + p_l y_l^1} = 1 - \text{Efficiency}^{-1}. \tag{5.7}$$

Because the right-hand side of Eq. (5.7) is positively related to efficiency (as there are two minus signs), we once more have a positive relationship between efficiency and profitability. But be careful. Under accounting prices, fully efficient firms break even and inefficient units operate at a loss. Indeed, if efficiency is one in formula (5.7), as is the case fully efficient firms, then the right-hand side is zero; hence, profit must be zero. If the efficiency is less than full, for example 3/4, then the right-hand side would be −1/3, creating a loss on the left-hand side.

It is not difficult to evaluate the expressions numerically, and I will illustrate this for the example of Fig. 5.3. Accounting prices can be computed using the zero

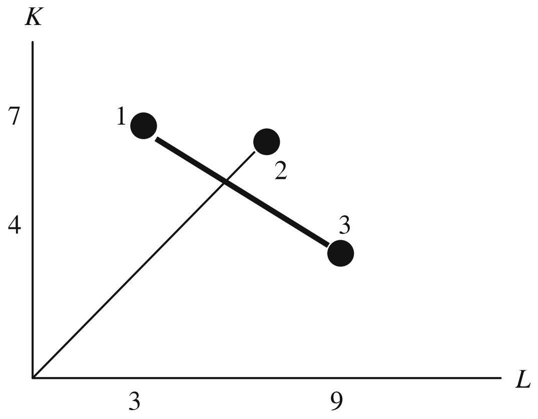


Fig. 5.3 An industry with two inputs and a single output. Firms *1* and *3* are efficient. Firm *2* is not. Yet, if labor is very expensive, input combination *3* is more costly than input combination *2*

profit conditions for the efficient firms. In Fig. 5.3, the accounting profit of firm 1 is $1 - (3w + 7r)$, where w is the wage rate and r the rental rate, while the accounting profit of firm 3 is $1 - (9w + 4r)$.[2] Setting these two profits equal to zero, we obtain two simple equations: $3w + 7r = 1$ and $9w + 4r = 1$. The solution is $w = 1/17$ and $r = 2/17$. These figures can be used to calculate the total cost of firm 2, hence its accounting profit (which will be negative) and efficiency, using formula (5.7). Let me carry out the calculation. Reading Fig. 5.8, firm 2's inputs are 7 units of labor and 6 of capital, and hence, its profit is $1 - (7/17 + 6 \times 2/17) = 1 - \mathit{Efficiency}^{-1}$, using formula (5.7). Solving, the efficiency of firm 2 amounts 17/19. In general, the numbers are given simply as output by the linear programming routine.

The crucial difference between market prices and accounting prices is that the former are observed and the latter are not. Market prices are exogenous, meaning that they are considered given, as *data*. Accounting prices are endogenous, meaning that they are derived from the data, not from price data, but from input and output data. It is instructive to see the difference in the simple single-input/single-output duopoly, where firm 1's data are input x^1, output y^1, prices $\underline{w}^1$ and $\underline{p}^1$, and firm 2's data are x^2, y^2, $\underline{w}^2$, and $\underline{p}^2$. Different price normalizations are allowed. If we stick to Sect. 5.2, Eq. (5.3), it is $py^1 = 1$, and the zero profit condition, $py^1 - wx^1 = 0$, yields $w = py^1/x^1 = 1/x^1$. Instead of the accounting prices $p = 1/y^1$, $w = 1/x^1$, we may use $p = 1$, $w = y^1/x^1$ as the relative prices are the same and that is all what matters. There are two ways to understand this. One is via result (6), equating efficiency with the revenue/cost *ratio*. This measure is clearly insensitive with respect to proportionate price changes. The other is to visualize a change in the scale of measurement. For example, if firm 1 produces 100 kg of rice, then the normalization condition $py^1 = 1$ reads $p = 1/100 = 1$ cent per kilogram. Now, we *could* choose this as a new currency unit, i.e., the cent instead of the dollar. Then, the value of output would not be 1, but 100 cents and the price would be 1 cent or 0.01 only.

The bottom line is that accounting prices must be *derived* from quantity data, in a way such that the most productive firms break even. Profitability implications of performance may be misleading when external, possibly distorted prices are used. For example, if the less productive firm, firm 2, commands a lower input price, $\underline{w}^2 < \underline{w}^1$, it may be equally profitable or even more so than firm 1, if the input price discount is strong. The question arises what to do in such a situation. Is it advantageous to stimulate the *productive*, firm 1, or the *profitable*, firm 2? In other words, where should one allocate the industry's resources?

The answer may vary with the setting, but a general observation is in order. *Both* firms would benefit from adopting the most productive technique. In the simple example, this would not change a thing to firm 1 (which already is the best practice firm), but it clearly would make firm 2 more profitable. Firm 2 earns a profit of $\underline{p}^2y^2 - \underline{w}^2x^2$, but replacement of output y^2 by potential output $(y^1/x^1)\,x^2$ would add to the revenue term and hence increase profit. Here, we recognize y^1/x^1 as the best

[2] Revenue is 1 because of the single output variant of price normalization condition (5.3), Sect. 5.2.

practice productivity. Alternatively, should the market not bear the additional output, the profit could be increased by cutting back input x^2 to what is necessary given the best practice technique, namely $(x^1/y^1)\ y^2$.[3] Here, we recognize x^1/y^1 as the minimal technical coefficient.

The productivity and the technical coefficient are each other's inverse, which is no surprise, because productivity is basically output per unit of input and a technical coefficient is input required per unit of output. The lesson of this example is that to improve performance, one must be on the look for the most *productive* practice, not the most profitable. This is the relevant rule of thumb even if the criterion is profit. In other words, the profit of a firm is enhanced by adopting best practice techniques, not by adopting the most profitable practices. In our duopoly example, the adoption of the technique of the most profitable firm (firm 2) would even be detrimental to the profit of firm 1!

This paradoxical relationship between productivity and profitability rests on the following fact: *One may copy techniques, but not prices.* Emerging economies rightly adopt Western production practices in manufacturing and the service sectors, because like everyone, they benefit from efficiency. Conversely, these Western plants do not copy their Eastern counterparts, even though they may be more profitable. The low wages prevailing in China and its Southern neighbors cannot be copied. They are reflections of conditions beyond business control, such as the endowments of nations. If a nation is well endowed with labor relative to other resources such as minerals, local wages will be low. This argument is valid even in the absence of exploitation.

5.4 Ranking

Ranking is a persuasive management tool that provides a sense of direction and infiltrates all corners of the information society. This section discusses the subtleties that surround this main application of benchmarking. The basic idea is to calculate the efficiencies of firms and to line them up between 0 and 100 %, but there is a complication. Efficiency is measured by the revenue/cost ratio at accounting prices —see Sect. 5.2, Eq. (5.6)—but these prices vary across firms. The accounting price of an input measures how much more output could be produced if an extra unit of the input was available. Now if an input is scarce in a firm, it acts as a bottleneck and, therefore, carries a high accounting price. Since the mix of inputs may differ across firms, an input may be relatively scarce somewhere and abundant elsewhere. This is why accounting price vary across units. If the industry is well organized, such differences are leveled. This observation will be clue to the measurement of industrial organization efficiency. The rule of thumb is as follows.

[3] The latter is indeed less than x^2 by assumption that firm 1 is more productive, $y^1/x^1 > y^2/x^2$.

Reallocate the excess resources of less efficient firms to an efficient firm where the accounting price is relatively high.

The rationale of this rule of thumb is that resources are best put to work where they are most productive and *accounting* prices are equal to marginal productivities. Market prices do not have the power to signal where resources are best put to work, simply because they are equal for different firms. On the product side, the relationship is the opposite.

Idiosyncratic accounting prices are higher for outputs which are produced relatively abundantly.

This result is perhaps paradoxical, because we tend to associate abundance with low prices. However, large-scale production drains resources and, therefore, is costly indeed. The negative relationship between quantity and price is a property of demand functions, whereas here we analyze the supply side of a firm. Then, the relationship is opposite indeed.

Recall that we have firms i with input vectors x^i and output vectors y^i. We calculate the maximally producible output ey^i, given the input x_i and the practices (x^i, y^i), see program (5.1), where symbol e stands for the expansion factor to be maximized. The dual Eq. (5.2) generates the accounting prices of the inputs, w_1, …, w_k, and of the outputs, p_1, …, p_l. (Here, k and l are the numbers of the inputs and the outputs, respectively.) The problem is that these prices are specific to the object we benchmark: firm I, because inputs may be scarce at some firms and abundant at others and outputs may be produced in costly volumes.

In the inequalities of program (5.1), the intensities θ_1, …, θ_I are the variables and the right-hand sides are firm specific and prompt the accounting prices to be idiosyncratic. Imagine that we have the power to improve the performance of firms not only by letting them adopt best practices, but also by reallocating their resources. The formal analysis involves the assessment of the overall efficiency of the industry by calculating how much more total output it could produce given its *total* input. Instead of benchmarking firms, we benchmark the entire industry. The benchmarking continues to be done on the same reference group of peers, i.e., on its own firms and *not* on others. This procedure amounts to replacement of the right-hand sides of constraints (5.1) by the *total* industry resources or inputs x_1, …, x_k and (potential) outputs ey_1, …, ey_k, where e is the expansion factor, as before. These total figures are defined by the following equations:

$$\begin{aligned} x_1 = x_1^1 + \cdots + x_1^I, \cdots, x_k = x_k^1 + \cdots + x_k^I, \\ y_1 = y_1^1 + \cdots + y_1^I, \cdots, y_l = y_l^1 + \cdots + y_l^I. \end{aligned} \tag{5.8}$$

The overall efficiency of the industry is the inverse of the expansion factor, e, where the latter is the solution to benchmarking program (1) with Eq. (5.8) used to modify the right-hand sides:

$$\begin{array}{l} \max_{\theta_1,\ldots,\theta_I,e\geq 0} e: \\ x_1^1\theta_1+\cdots+x_1^I\theta_I\leq x_1,\cdots,x_k^1\theta_1+\cdots+x_k^I\theta_I\leq x_k \\ y_1^1\theta_1+\cdots+y_1^I\theta_I\geq y_1 e,\cdots,y_l^1\theta_1+\cdots+y_l^I\theta_I\geq y_l e \end{array} \tag{5.9}$$

As we did in Sect. 5.2 for firm i, associate input and output accounting prices w_1, …, w_k and p_1, …, p_l with benchmarking program (9). The attractive property of these prices is that they measure the marginal productivity of inputs and outputs to the industry as a whole. The accounting prices thus constructed are independent of the firm under consideration.

The resolution of the weighting problem rests on the replacement of firm's scarcities and abundances by their overall counterparts. For a single-output industry, the accounting prices of a firm with a representative mix of inputs are the ones which reflect the values to the industry as a whole, in the sense of marginal productivities. An analogous observation can be made for a single-input/multiple-output industry. *The performance weights generated by the benchmarking program of a firm with a representative mix of scores agree with the optimal ones (relevant to the industry as a whole).* There is no need to identify the firm of which the accounting prices can be used to measure and compare the performance of all units. It suffices to solve the industry's benchmarking program, (9). The output contains the shadow prices.

Having settled the issue of weighting performance dimensions, let us now tackle the issue of ranking. We employ the input and output accounting prices w_1, …, w_k and p_1, …, p_l associated with benchmarking program (9). The efficiency of firm i is given by the revenue/cost ratio (6). The weights are independent of the firm! The firms are considered machines which transform inputs into outputs. The outputs are aggregated using the weights p_1, …, p_l and the inputs with the weights w_1, …, w_k. Expression (5.6) measures efficiency as aggregated output per unit of aggregated input. The theory of Sect. 5.3 applies in particular the observation that efficiency is a measure between zero and one. The reason is that the duality analysis—see Eq. (5.4) by which the value of the outputs is less than or equal to the value of the inputs—happens to be *independent* of the object that is benchmarked (i.e., firm i in Sect. 5.2 or the entire industry in the present section).

5.5 Economies of Scale

There are four types of returns to scale, namely constant returns, decreasing returns, increasing returns, and variable returns to scale. The principles are easiest understood for a single-input/single-output production unit, where the input and output are denoted by x and y, respectively.

In Fig. 5.4, the first panel represents the case of *constant returns to scale*. An increase in the input quantity yields a proportionate increase in the output quantity. The second panel represents the case of *decreasing returns to scale*. Here, the

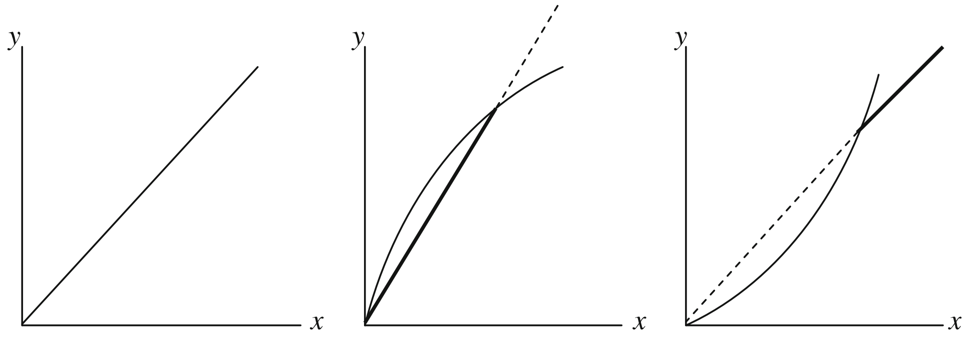


Fig. 5.4 Constant, decreasing, and increasing returns to scale defined. In each panel, input is along the horizontal axis and output along the vertical. The unbroken lines are below the production function. The thin dashed lines are above the production function and, therefore, not feasible

returns of additional input are less than proportionate, for example, 1 % of extra input yields only 0.9 % of extra output. The third panel represents the case of *increasing returns to scale*, where the returns become more than proportionate.

If there are constant returns to scale—the left panel—any feasible activity, represented by an input–output combination (x, y) on the graph (or under it, but that would be wasteful), can be run with *any* nonnegative intensity, θ. If (x, y) is feasible, then so are (θx, θy) with $\theta \geq 0$.[4] In other words, if a point is feasible, then so is *any* other point on the *half line* through that point and the origin. However, if there are decreasing returns to scale—as in the central panel of Fig. 5.4—any feasible activity can be run with *lower* intensity only. In other words, if (x, y) is feasible, then so are (θx, θy) with $0 \leq \theta \leq 1$. If a point is feasible, then so is any other point on the line unbroken *segment* connecting that point with the origin. Finally, if there are increasing returns to scale—as in the right panel—any feasible activity can be run with *higher* intensity only. If (x, y) is feasible, then so are (θx, θy) with $\theta \geq 1$. In other words, if a point is feasible, then so is any other point on the unbroken *outer half line* through the point, away from the origin. These observations are summarized in Table 5.2.

Now let me turn to the case of variable returns. First, I review some basic production theory. A flexible form for a production function is the S-shaped function. It features first increasing and eventually decreasing returns to scale, see Fig. 5.5.

In Fig. 5.5, a minimum quantity of input, F, is required to produce any positive amount of output, however little. This is called the fixed cost or overhead. Now, if output is increased, the fixed cost can be spread among more units and this causes the returns to scale to be initially increasing. The effect peters out though. For big corporations, overhead costs—however sizable in an absolute sense—become a small percentage of total cost, and another scale effect sets in, namely that of bottlenecks. Some inputs are just very hard to increase, think of land, and eventually limit output as the variable inputs are increased. At some intermediate

[4] (θx, θy) can be denoted briefly by θ(x, y).

Table 5.2 Returns to scale and feasible intensities

Returns to scale	Feasible intensities	Geometry	Panel in Fig. 5.1
Constant	$\theta \geq 0$	Half line	Left
Decreasing	$0 \leq \theta \leq 1$	Line segment	Central
Increasing	$\theta \geq 1$	Outer half line	Right

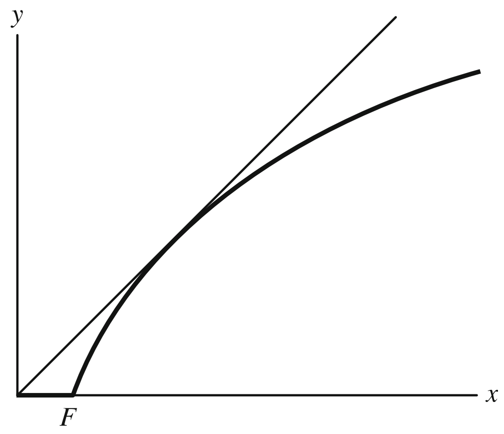


Fig. 5.5 An S-shaped production function. Input is along the *horizontal* axis and output along the *vertical*. There is a fixed cost (F)

level, the two scale effects balance and productivity (output–input ratio y/x) is maximal. This is where the line to the origin is steepest, see the straight line in Fig. 5.5. To the left of this point of tangency, the returns to scale are increasing and to the right decreasing. Productivity may be maximal in a *region*, see Fig. 5.6.

With a little imagination, one recognizes an S-shape in Figs. 5.5 and 5.6. Better known is the so-called U-shaped average cost, the other side of the coin. The reason is simple: Average cost is determined by input per unit of output, which is x/y or inverse productivity. Since productivity is initially increasing and eventually decreasing, average cost is initially decreasing and eventually increasing, hence U-shaped. The U-shaped average cost associated with the production function of Fig. 5.6 has a flat bottom.

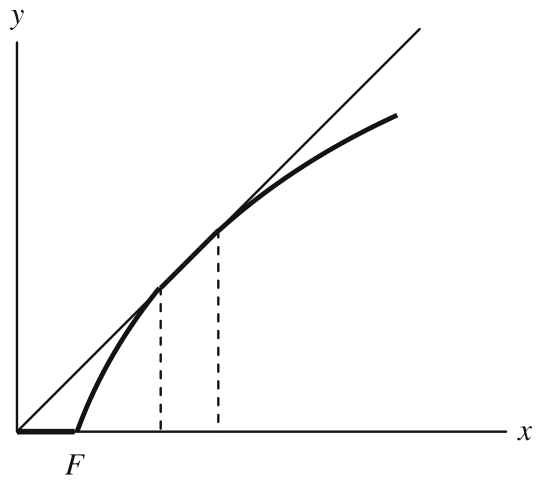


Fig. 5.6 Another S-shaped production function. Input is along the *horizontal* axis and output along the *vertical*. In the region between the *dashed lines*, productivity is maximal

The S-shaped production function is realistic, because it combines setup costs with bottleneck effects. It is also flexible, because the point of maximal productivity may be reached when input is arbitrarily small—in which case the returns to scale are decreasing right away—or when input is arbitrarily large—in which case the returns to scale remain increasing for all relevant levels of activity. In other words, the S-shaped production function *encompasses* the cases of decreasing and increasing returns. In this sense, it is quite general and it is desirable to have a counterpart in a multi-input multi-output framework for efficiency measurement.

The counterpart exists and is called *variable returns to scale*. The model is due to Afriat (1972), and Førsund and Hjalmarsson (1974) and has been launched in a DEA setting by Färe et al. (1983) and Banker et al. (1984). The idea is that fixed costs cannot be dissolved by running firms, including their inputs, at small intensities. One may combine firms, but the level of operation must remain the same. Formally, *the sum of the intensities must be one*. A simple example illuminates, see Fig. 5.7.

Figure 5.7 features a small firm (the lower dot) and a big one (the upper dot). It is drawn such that the small firm is more productive (more output per input, the line to the origin is steeper), but that does not matter and could have been the other way round. The tenet of DEA with variable returns is that any weighted average of the observed firms is feasible. These weighted averages are represented by the line segment connecting the two dots. It is assumed that it is impossible to run single activities at lower intensity. If allowed, variable returns to scale would degenerate into decreasing returns to scale. The variable returns to scale model also excludes the possibility to run activities at higher intensities. Doing so would take us back to the case of increasing returns to scale.

An implicit but important assumption of DEA with variable returns to scale is that the hypothetical firm representing inactivity—with zero input and zero output—is ruled out. The reason is simple. If it were allowed, any firm could dissolve its fixed cost by averaging out with the inactivity point (the origin in Fig. 5.7), and this trick would take us back to the small-is-beautiful world of decreasing returns to scale, such as Fig. 5.1, middle panel. The bottom line is that the smallest observed fixed cost is accepted as inescapable.

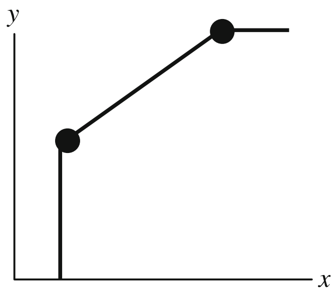


Fig. 5.7 S-shaped production observed. Input is along the *horizontal* axis and output along the *vertical*. The fixed cost cannot be escaped

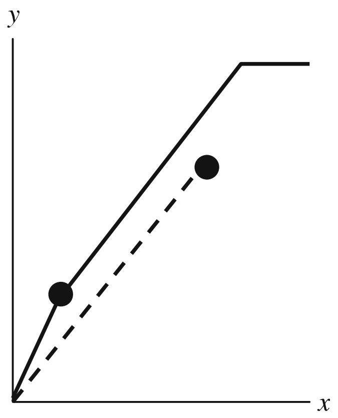


Fig. 5.8 Same data as Fig. 5.7, but assuming decreasing returns. Input is along the *horizontal* axis and output along the *vertical*. The two points sum to the unidentified

The relationship between the analysis of returns to scale and data envelopment involves a subtle distinction. The issue can be explained in the context of the simple Fig. 5.7. Perhaps the most natural assumption in this example would be that of decreasing returns to scale, for the small unit, has a greater output/input ratio than the big unit. The consequent production function is depicted in Fig. 5.8.

Let me explain Fig. 5.8. Under the assumption of decreasing returns to scale, any activity can be run at a lower scale as well. This explains the line segment connecting the origin with the small unit and also the one connecting the origin with the big unit. At small levels of input, the maximum level of output is determined by the output/input ratio of the small unit, as it is greater than the output/input ratio of the big unit. Now suppose we command a level of input slightly above the size of the small unit. What is the maximum amount of producible amount? Well, first employ the first unit up to capacity, as it is the more productive. The remaining available inputs are employed in the second unit and increase output beyond the first data point in Fig. 5.8, at a rate determined by the productivity of the second data point, which is the slope of the dashed chord. We translate that chord from the origin to the first data point and thus continue the production.

Figure 5.7 shows the tricky difference between scale economies in economic theory and DEA. For example, imagine the total input is reduced to the input of the big unit: An accident kills the workers in the small business. Then, it would be optimal to relocate workers from the big business to the small business, which is more productive in Fig. 5.8. This would increase output. Hence, the output of the big business is not the maximum producible output. In other words, the second data point resides *within* the production possibility frontier and the frontier is *not* the closest envelopment of the data.

Although it is not clear if it is desirable, the discrepancy between returns to scale assumptions and data envelopment can be resolved. If we reason like many free market economists do, we would argue that if the big business could produce more it would produce more and, therefore, it better resides on the production possibility frontier. This mildly dogmatic reasoning can be accommodated by the following modification, which is motivated by the analysis of variable returns to scale. Instead of assuming that *each* firm's intensity is less than unity (Table 5.1, decreasing returns), assume that the *total* intensity is less than unity: $\theta_1 + \theta_2 \leq 1$. (This is for two units, as in Fig. 5.8. The extension to more units is straightforward.)

In DEA, enveloping the data, output expansion beyond the full utilization of the productive, small business is possible only by simultaneously shrinking the utilization of that same business. This process continues gradually until the next unit is fully utilized and the first, small unit no longer. Hence, the connection between the two data points in Fig. 5.9. The consequent frontier is below the one of ordinary decreasing returns, depicted in Fig. 5.8.

The concept of increasing returns to scale can be modified similarly. Instead of assuming that *each* firm's intensity is greater than unity (Table 5.2, increasing returns), we assume that the *total* intensity is greater than unity.

All the forms of returns to scale we have encountered are applicable to organizations with multiple inputs and outputs. It is a bit messy in the framework of production functions, see, e.g., Baumol (1977), but in the application to benchmarking, the analysis becomes pleasantly crisp. The different returns to scale cases can be described by alternative restrictions on the intensities with which best practices are run to assess the potential output of a firm. Table 5.3 collects the encountered returns to scale cases.

Roughly speaking, Table 5.3 shows that constant/decreasing/increasing/variable returns to scale are a matter of feasible intensities to be any/below unity/above unity/unity. Here, "intensities" are individual intensities in the economics literature and total intensities in the DEA literature. Under constant returns to scale, there is full consistency between the two approaches.

Under decreasing returns to scale, there is a one-way consistency. Here, DEA feasible intensities are also feasible in the sense of economics. This means that the feasible production possibility set in DEA is smaller than in economics. Hence, the DEA frontier will be closer than the economics frontier. *Hence, under decreasing returns, DEA overestimates efficiency relative to the economic model.*

With increasing returns to scale, there is also consistency, but the other way. Now intensities feasible in the sense of economics are also feasible in the sense of DEA decreasing returns. This means that the feasible production possibility set in

Fig. 5.9 Enveloping the data of Fig. 5.7 or 5.8. Input is along the *horizontal* axis and output along the *vertical*

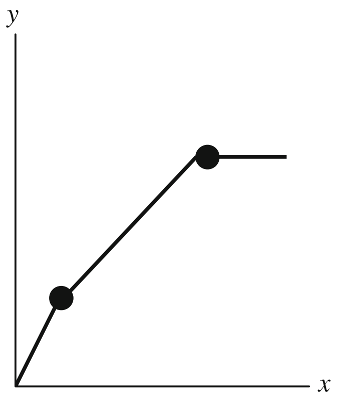


Table 5.3 Returns to scale in economics and DEA

Returns to scale	Feasible intensities		
	Economics	Relationship	DEA
Constant	$\theta_i \geq 0$	=	$\theta_i \geq 0$
Decreasing	$0 \leq \theta_i \leq 1$	<=	$\theta_i \geq 0$ and $\sum\theta_i \leq 1$
Increasing	$\theta_i = 0$ or ≥ 1	=>	$\theta_i \geq 0$ and $\sum\theta_i \geq 1$
Variable	Closed interval	<=	$\theta_i \geq 0$ and $\sum\theta_i = 1$

DEA is bigger than in economics. Hence, the DEA frontier will be farther out than the economics frontier. *Hence, under increasing returns, DEA underestimates efficiency relative to the economic model.*

Variable returns to scale are locally increasing or decreasing, at least at frontier points. Intensities may increase or decrease, respectively. The feasible set is $[1, \theta_{\max}]$ or $[\theta_{\min}, 1]$, respectively. For nonfrontier points, it is $[\theta_{\min}, \theta_{\max}]$. It is not possible to specify these intervals without external information. Hence, in this paper, only the DEA variant of variable returns to scale will be considered. The DEA variant is restrictive. Observations are considered efficient, except when they are dominated by convex combinations of others. *Variable returns to scale in the sense of DEA are an instance of variable returns to scale in the economics sense.*

Following our discussion, summarized in Table 5.3, we can limit the intensities of activities. This amounts to the addition of alternative constraints to program (1). In the case of *decreasing* returns to scale, the condition that an intensity of a benchmark (or the sum of these intensities if we take the DEA tack) must be between 0 and 1 automatically includes the possibility of ignoring bad examples (by setting their intensities equal to corner value zero). This straightforward extension preserves the linearity of program (1). *Increasing* returns to scale, however, includes the possibility $\theta = 0$ as a distinct event, creating a nonconvexity. Under *variable* returns to scale, this complication is happily dissolved again. The set of admissible intensities is defined by the simple summing up condition given in Table 5.3, which admits smooth transitions between potential benchmarks. The variable returns to scale model also inherits the property of S-shaped production function that it encompasses the cases of decreasing and increasing returns. Related to this flexibility is the better fit to the data. As there is empirical support for S-shaped production functions (or U-shaped average costs), the variable returns to scale model of benchmarking outperforms its counterparts. The variable returns to scale model can be used to locate firms, placing them in the initial region of increasing returns, the intermediate efficient region of constant returns, or the eventual region of decreasing returns.

5.6 Alternative Efficiency Measures

In the *decreasing returns to scale* model, we add the *capacity* constraints, θ_1, …, $\theta_I \leq 1$ (the economics approach) or $\theta_1 + \cdots + \theta_I \leq 1$ (DEA). Denoting the shadow prices of the separate capacity constraints by τ_1, …, $\tau_I \geq 0$, the dual constraints associated with the intensity variables become as follows:

$$py^1 \leq wx^1 + \tau_1, \ldots, py^I \leq wx^I + \tau_I \tag{5.10}$$

The dual constraint associated with the final variable (e) is the price normalization constraint (3). If the constraints are pooled (the DEA approach), we have *commonality* of the shadow prices, $\tau_1 = \cdots = \tau_I$. In either case, the difference with the basic model is that decreasing returns to scale generate *profits*. The profits are determined by the shadow prices of the intensity constraints (θ_1, …, $\theta_I \leq 1$).

By the phenomenon of complementary slackness, positive profit implies that the unitary intensity constraint must be binding. This implies that its nonnegativity constraint is not binding. This implies—once more invoking complementary slackness—that there is no slack in the shadow prices of the nonnegativity constraints. By the theory of linear programming, this means that the inequality for such a firm in Eq. (5.10) reduces to an equality. In short, profit $\tau_i > 0$ implies equality for unit i in (10). Conversely—the logical negation of the last sentence—inequality for a unit in (10) must yield no profit, $\tau_i = 0$. Incorporating these insights, Eq. (5.10) may be rewritten as follows:

$$py^j = wx^j + \tau_j \quad \text{or} \quad py^j < wx^j \tag{5.11}$$

The left-hand side of Eq. (5.11) are the benchmarks for the decision unit we assess. On the right-hand side are units which have positive shadow prices of their nonnegativity constraints. By the phenomenon of complementary slackness, these units have binding nonnegativity constraints, and hence are inactive and, therefore, constitute no benchmark.[5]

Efficiency is the inverse of the expansion factor, e. By the main theorem of linear programming, e is the value of the bounds in program (1) and the added constraint from Table 5.3. The former is x^i, applied to the inputs. The latter are 1, applied to the intensities (individual in the economic frame or total in DEA). Hence, $e = wx^i + \tau_1 + \cdots + \tau_I$, or, invoking price normalization constraint (3),

$$\text{Efficiency} = py^i/(wx^i + \tau_1 + \cdots + \tau_I) \tag{5.12}$$

[5] It is possible that $\tau_j = 0$ in Eq. (5.11). This corresponds to a benchmark for which neither the nonnegativity constraint (as signaled by the equality) nor the capacity constraint (as signaled by the zero value of its shadow price, τ_j) is binding, i.e., the unit is hovering at partial capacity: $0 < \theta_j < 1$.

Remark on formula (5.12). In the DEA, variant profit $\tau_1 + \cdots + \tau_I$ is replaced by just the single (common) profit. Since this profit depends on the unit we benchmark, i, I denote it by τ^i. This confirms Table 5.3 and the ensuing discussion by which the DEA efficiency level is greater. Also, if the unit is its own benchmark, the left-hand side holds in Eq. (5.11), or $py^i = wx^i + \tau^i$, so that Efficiency $= py^i/(wx^i + \tau^i) = 1$, as should be.

In the *increasing returns to scale* model, the benchmarks can be determined in a similar way, albeit that we must now consider the many discrete possibilities mentioned before. Suppose we have done this and let $\underline{I}$ be the subset of *active* units j in program (1) (augmented with an increasing returns condition of Table 5.3), for which θ_j are positive. The other, *inactive*, units do not contribute to the determination of the maximal producible output and, therefore, may be ignored. I relabel the units such that the active ones are listed up front, $i = 1, \ldots, \underline{I}$. The inactive units are $i = \underline{I} + 1, \ldots, I$. It follows that program (1) may be replaced by,

$$\max_{\theta_1,\ldots,\theta_{\underline{I}} \geq 1, e} e{:}x^1\theta_1 + \cdots + x^{\underline{I}}\theta_{\underline{I}} \leq x^i, y^1\theta_1 + \cdots + y^{\underline{I}}\theta_{\underline{I}} \geq y^i e \tag{5.13}$$

Program (5.13) modifies program (5.1) in two ways: The value of the lower bounds, 0, becomes 1, and the replacement I becomes $\underline{I}$. Hence, the dual equation is the same as in the basic case, (5.10) without τ's, and with $\underline{I}$ instead of I:

$$py^1 \leq wx^1, \ldots, \quad py^{\underline{I}} \leq wx^{\underline{I}} \tag{5.14}$$

If an inequality in (5.14) is strict (<), the slack is positive. Since the slack is the shadow price of the ≥1 constraint, the intensity equals 1 (by the phenomenon of complementary slackness). By the main theorem of linear programming, e is the value of the bound in program (1), x^i, and the added constraint from Table 5.3, number(s) 1 (individual in the economic framework or total in DEA). The added constraint is flipped compared to the decreasing returns case. Hence, the shadow prices are $\tau_1, \ldots, \tau_I \leq 0$: profits turn *losses*. With this modification, *efficiency formula* (5.12) *and the remark on the latter remain valid.*

Last but not least, let me address the case of *variable returns to scale*. The analysis follows the decreasing returns to scale case, with two tricks added. The first trick is that we now add a *single* capacity constraint to the basic model: $\theta_1 + \cdots + \theta_I = 1$. I denote the shadow price of this constraint by τ. Since the constraint is an equality, this shadow price is now *unsigned* and dual constraint (10) becomes as follows:

$$py^1 \leq wx^1 + \tau, \; \ldots, \; py^I \leq wx^I + \tau \tag{5.15}$$

The main difference of Eq. (5.15) compared to the basic model is that with variable returns to scale, accounting prices admit profits *or* losses. Moreover, compared to the case of decreasing returns, the accounting prices are now such that the profit is uniform across the benchmarks. Efficiency remains given by the DEA variant of Eq. (5.12), with $\tau_1 + \cdots + \tau_I$ replaced by τ^i (see the remark):

$$\text{Efficiency} = py^i/(wx^i + \tau^i) \tag{5.16}$$

The variable returns to scale model seems the most popular benchmarking tool. It envelops the data closely and hence reduces the estimates of inefficiency, because the latter is measured by the gap between an observation and the frontier representing potential output. This reduction is considered not bad, because it offsets a shortcoming of DEA, namely its tendency to overestimate inefficiency.[6]

Scale returns may vary by nature (constant/decreasing/increasing/variable) and by frame of reference (economic/DEA). The alternatives are represented by alternative constraints on the feasible intensities (Table 5.3). These translate in alternative shadow prices. In view of the (non)convexity of the different cases, it is perhaps surprising that a single efficiency formula, (12), applies to *all*. The profits are nonnegative (decreasing returns), nonpositive (increasing returns), or either (variable returns). Moreover, the profits are firm specific in the economics framework and uniform in DEA.

A central concept in industrial organization is the industry production function. The concept is best introduced for a single-input/single-output industry. Let the production function be F. Examples are depicted in Fig. 5.1. With I firms, the producible output is $y = F_I(x) = \max F(x^1) + \cdots + F(x^I)$ subject to $x^1 + \cdots + x^I = x$. With free entry (and exit), the industry production function is defined by $y = \max_I F_I(x)$.

With *constant* returns to scale, the number of firms is irrelevant and the program that defines the potential output of the firm, (1), also specifies the industry output.

Ever since McKenzie (1959), it is known that if the production function features *decreasing* returns to scale, the free-entry industry production function has constant returns to scale, with the optimal $I = \infty$. Because of this incompatibility, it is common to preserve the number of firms, I. Let us return to the example of two firms (discussed in Sect. 5.3). Combining Figs. 5.8 and 5.9 and adding the total input/total output bundle, we get Fig. 5.10.

What is the value of the industry production function at the level of total input? Well, it depends on the framework. The upper graph represents the economics production function and the lower graph the DEA production function or envelop. To determine the value of the industry production function, in either case, we may reallocate the inputs of the two firms, given by x^1 and x^2, to $x^{1\prime}$ and $x^{2\prime}$, respectively, such that total input is preserved—$x^{1\prime} + x^{2\prime} = x^1 + x^2$—and apply program (1) to each part, $x^{1'}$ and $x^{2\prime}$. The "new" firm, now commanding $x^{1'}$, runs the

[6] The problem is that DEA is sensitive with respect to errors of measurement, particularly of best practice observations. Overstatement of output or understatement of input may falsely identify decision-making units as benchmarks, and as a consequence, throw back the other decision-making units. The latter are thus suggested being relatively inefficient, but it is a fluke in the data.

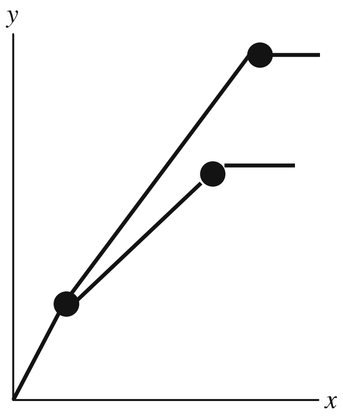


Fig. 5.10 Same data as Fig. 5.8, total input/total output added. Input is along the *horizontal* axis and output along the *vertical* axis

activities (represented by the two firms) with intensities θ_1^1, θ_2^1. The new firm now commanding $x^{2\prime}$ runs the activities with intensities θ_1^2, θ_2^2. By linearity, this is equivalent to running the two activities with intensities $\theta_1^1 + \theta_1^2, \theta_2^1 + \theta_2^2$. Call these θ_1, θ_2. The question of feasibility (not using more than total input) can be addressed by transforming the intensity constraints defining alternative returns to scale (on the superscripted θ's) from the level of the firms to the level of the industry (the nonsuperscripted θ's).

In *DEA,* the decreasing returns to scale constraints are $\theta_1^1 + \theta_2^1 \leq 1$ and $\theta_1^2 + \theta_2^2 \leq 1$. The *industry constraint* becomes $\theta_1 + \theta_2 \leq 2 = I$. Put numbers on Fig. 5.10, say inputs 1 and 3 and outputs 2 and 4, respectively. Consider a departure from the observed intensities, $\theta_1 = \theta_2 = 1$, to say $\theta_1 = 1.1$ and $\theta_2 = 0.9$. The inputs change by 0.1 − 0.1 × 3, which is feasible. The outputs change by 0.2–0.4, which is suboptimal. We may also consider a contraction of θ_1, but the released input will be less productive in the other activity. Hence, $\theta_1 = \theta_2 = 1$ determines the value of the 2 firms industry production function under decreasing returns to scale of the DEA variety.

In the *economics* approach, the returns to scale constraints are $\theta_1^1, \theta_2^1 \leq 1$ and $\theta_1^2, \theta_2^2 \leq 1$. The *industry constraint* becomes $\theta_1, \theta_2 \leq 2 = I$. In Fig. 5.9, it is optimal to maximize the small, more productive process: $\theta_1 = 2$. The residual input (net of the endowments), 1 + 3 − 2 × 1 = 2, determines the activity level of the other process, hence $\theta_2 = 2/3$.

One rule holds for either decreasing returns to scale approach (see Table 5.3): *The transition to the industry production function is achieved by replacement of 1 by I as the intensity constraint.* This procedure reflects the pooling of the inputs —$x^{1\prime} + x^{2\prime} = x^1 + x^2$.

The case of *increasing* returns to scale is also defined by separate intensity constraints (economics) or a pooled one (DEA). The difference is that the inequalities are flipped from ≤ to ≥1. Hence, *the same procedure holds for increasing returns to scale*.

As defended in Sect. 5.3, of *variable* returns to scale, only the DEA variant is considered. It is defined by the condition that for each firm activity, intensities sum to unity. Repeating the reallocation argument introduced in our discussion of decreasing returns, this translates into the condition that intensities sum to the number of firms. Hence, *the same procedure holds for variable returns to scale*.

The final step is to allow for free entry, varying the number of firms. This is problematic, particularly when there are variable returns to scale. The problem is well known for single-input/single-output industries with an S-shaped production function, which translates into U-shaped average costs. Output is maximized when firms minimize unit costs. The optimal number of firms is output divided by the output produced at minimum unit cost, but this number need not be an integer. This integer problem diminishes as the optimal number becomes large, and even disappears completely at some point if average costs are flat-bottomed (minimized at an interval of outputs instead of a single one). There are bounds for the magnitude of this problem, and they may be reestablished in the present DEA context. If the problem is ignored, admitting real values of the number of firms, I, it can be shown that the industry intensity constraints reduce to nonnegativity constraints, even in all three cases of scale economies. If the problem is not ignored, the optimal integer number of firms must be determined, for every total input/total output combination, by solving the following variable returns to scale industry program for every integer number of firms, I'.

$$\max_{\theta_1,\ldots,\theta_I \geq 0, \theta_1+\ldots+\theta_I=I', e} e: \\ x^1\theta_1 + \cdots + x^I\theta_I \leq x^1 + \cdots + x^I, y^1\theta_1 + \cdots + y^I\theta_I \geq (y^1 + \cdots + y^I)e \tag{5.17}$$

The difference vis-à-vis the efficiency program for a firm is in the coefficient of the expansion factor, e, and in the input and intensity bounds. That coefficient change merely affects the dual-price normalization constraint. The bounds changes merely affect the value of the primal or dual program. Dual constraint (15) is not changed. Efficiency Eq. (5.16) becomes Eq. (5.18).

$$\text{Efficiency} = p(y^1 + \cdots + y^I)/[w(x^1 + \cdots + x^I) + \tau I] \tag{5.18}$$

If the integer problem is not important, $\tau = 0$ in Eq. (5.18).

5.7 Firms and Industrial Organization Efficiencies

We have seen how the measurement of efficiency can be transplanted from the firm to the industry, under alternative returns to scale assumptions. The industry efficiency, (18), is less than the market share weighted harmonic mean of the firm efficiencies, (16), and the ratio of the two is the efficiency of the industrial organization, as will be detailed in this section. Point of departure is industry program (9). Potential output is determined by not only letting firms adopt best practices, but also by relocating resources between them. In other words, potential industry output exceeds the sum of the potential firm outputs. Since the gap between observed output and potential output measures inefficiency, it follows that the efficiency of the industry is less than what one would expect on the basis of the firm efficiencies. More precisely, ten Raa (2011) has shown that $1/e \leq 1/(s^1e^1 + \cdots + s^Ie^I)$, where s^i are the market shares of the firms, $py^i/p(y^1 + \cdots + y^I)$, evaluated at the shadow prices of program (9) (the Lagrange multipliers of the output constraints). The right-hand side of this inequality is the harmonic mean of the firm efficiencies, $1/e^1, \ldots, 1/e^I$. If, for example, the harmonic mean of the firm efficiencies is 80 %, but the industry efficiency is 60 %, the allocation of the resources between the firms is such that full firm efficiency attains only 60/80 = 75 % of full industry efficiency and we say that the efficiency of the industrial organization is 75 %. Formally, *industrial organization efficiency* is defined by the ratio of industry efficiency to mean firm efficiency: $\varepsilon^{IO} = (s^1e^1 + \cdots + s^Ie^I)/e$.

Full efficiency of an industry requires that all firms are efficient and that the industrial organization is efficient. ten Raa (2011) has shown that a necessary and sufficient condition is that the industrial organization is supportable (Sharkey and Telser 1978), meaning invulnerable to entry. This one expects in a contestable market (Baumol et al. 1982).

An immediate consequence of the definition of industrial organization efficiency is that industry efficiency is the product of mean firm efficiency and industrial organization efficiency.

We commingle efficiency change and technical change into productivity growth and bring in industrial organization change. For a solid conceptual foundation, consider the structure of an efficiency program. Efficiency is the inverse expansion factor, $\varepsilon = 1/e$. It is a function of the parameters in the efficiency program, such as (1). The right-hand sides of the constraints feature the inputs and outputs of the unit of which the efficiency is assessed, (x^i, y^i). The left-hand sides feature the inputs and outputs of all firms, which we write formally by (X, Y), where X is the matrix of all inputs (across firms) and Y the matrix of all outputs.

The efficiency of firm i is a function of its (own) input/output vectors and of the family of all these vectors (representing the industry): $\varepsilon^i = 1/e^i = f((x^i, y^i), (X, Y))$, where function f summarizes the efficiency program, presenting the inverse of the expansion factor as a function of the program parameters. The data, $((x^i, y^i), (X, Y))$, are a function of time and, therefore, so is efficiency. *Efficiency change* is defined by the percentage expression $\text{EC} = (d\varepsilon^i/dt)/\varepsilon^i$. Total differentiation of efficiency function f yields the following equation:

$$\text{EC} = (\frac{\partial f}{\partial x^i}\frac{\partial x^i}{\partial t} + \frac{\partial f}{\partial y^i}\frac{\partial y^i}{\partial t})/f + (\frac{\partial f}{\partial X}\frac{\partial X}{\partial t} + \frac{\partial f}{\partial Y}\frac{\partial Y}{\partial t})/f \tag{5.19}$$

The first term on the right-hand side of Eq. (5.19) measures the contribution of the firm, (x^i, y^i), by contracting its inputs or expanding its outputs, in short by improving its output/input ratio. Hence, this term defines *productivity growth*, $\text{PG} = \left(\frac{\partial f}{\partial x^i}\frac{\partial x^i}{\partial t} + \frac{\partial f}{\partial y^i}\frac{\partial y^i}{\partial t}\right)/f$. In the second term on the right-hand side of Eq. (5.19), the firm data are fixed, but the environment changes. Now, we must reason carefully. If there is technical progress, the efficiency of the firm will be *lower* in the next point of time, as its potential output will be greater under the new technology. Hence, the effect of the second argument measures *technical change*, but with a minus sign: $T = -\left(\frac{\partial f}{\partial X}\frac{\partial X}{\partial t} + \frac{\partial f}{\partial Y}\frac{\partial Y}{\partial t}\right)/f$. If we rearrange the terms in Eq. (5.19), we conclude that PG = EC + TC: *Productivity growth is the sum of efficiency change and technical change*.

All these concepts have discrete time variants, of which Malmquist indices have attractive theoretical properties. For productivity growth, the derivation is as follows. Since the derivative of ln *f* is 1/*f*, $\mathrm{PG} = \frac{\partial \ln f}{\partial x^i}\frac{\partial x^i}{\partial t} + \frac{\partial \ln f}{\partial y^i}\frac{\partial y^i}{\partial t}$, which in turn may be approximated by $\ln f(x^i_{t+1}, y^i_{t+1}, \cdot, \cdot) - \ln f(x^i_t, y^i_t,) = \ln \frac{f(x^i_{t+1}, y^i_{t+1}, \cdot, \cdot)}{f(x^i_t, y^i_t,)}$. (Subscripts *t* and *t* + 1 represent points in time, not components.) The dots represent the environment, (*X*, *Y*), evaluating it in times *t* and *t* + 1 and taking the average, $\mathrm{PG} = \ln \sqrt{\frac{f(x^i_{t+1}, y^i_{t+1}, X_t, Y_t)}{f(x^i_t, y^i_t, X_t, Y_t)} \frac{f(x^i_{t+1}, y^i_{t+1}, X_{t+1}, Y_{t+1})}{f(x^i_t, y^i_t, X_{t+1}, Y_{t+1})}}$, which can be approximated by $\sqrt{\frac{f(x^i_{t+1}, y^i_{t+1}, X_t, Y_t)}{f(x^i_t, y^i_t, X_t, Y_t)} \frac{f(x^i_{t+1}, y^i_{t+1}, X_{t+1}, Y_{t+1})}{f(x^i_t, y^i_t, X_{t+1}, Y_{t+1})}} - 1$. It is tradition not to subtract the one, so a *Malmquist productivity index* is simply defined by the geometric mean of productivity growth rates measured against the backdrops of times *t* and *t* + 1, $M_t = \sqrt{\frac{f(x^i_{t+1}, y^i_{t+1}, X_t, Y_t)}{f(x^i_t, y^i_t, X_t, Y_t)} \frac{f(x^i_{t+1}, y^i_{t+1}, X_{t+1}, Y_{t+1})}{f(x^i_t, y^i_t, X_{t+1}, Y_{t+1})}}$, and a Malmquist index of say 1.02 thus represents 2 % growth. To evaluate a Malmquist productivity index, one must determine four values of function *f* and hence solve four linear programs.

If we apply this analysis to an industry instead of a firm *i*, we trace the effects in ε = *f*((*x*, *y*), (*X*, *Y*)) and get the same decomposition, but since industry efficiency change is the product of mean firm efficiency and industrial organization efficiency, we obtain that *industry productivity growth equals industrial organization efficiency change plus mean firm efficiency change plus technical change*. ten Raa (2011) details the straightforward but tedious decomposition, with the result

$$M_t = \frac{\varepsilon^{IO}_{t+1}}{\varepsilon^{IO}_t} \times \frac{\sum s^i_t / f((x^i_t, y^i_t), (X_t, Y_t))}{\sum s^i_{t+1} / f((x^i_{t+1}, y^i_{t+1}), (X_{t+1}, Y_{t+1}))} \times \sqrt{\frac{f((x_t, y_t), (X_t, Y_t))}{f((x_t, y_t), (X_{t+1}, Y_{t+1}))} \times \frac{f((x_{t+1}, y_{t+1}), (X_t, Y_t))}{f((x_{t+1}, y_{t+1}), (X_{t+1}, Y_{t+1}))}}.$$

Symbols without superscripts represent industry totals, summed inputs or outputs, with all summations over firms *i* = 1, …, *I*. The first quotient is the index for industrial organization efficiency change, the second for mean firm efficiency change, and the third for technical change.

So far we have remained silent about entry and exit; the number of firms, I, has been constant. The idea that industry efficiency is a combination of firm efficiencies and organization efficiency is an aggregation result for firms that can be extended to groups of firms. I model entry as an extension of the number of firms and exit as the transition to dormant, $x^i = 0$, $y^i = 0$. The number of firms becomes $I + E$, where the first term represents the incumbents and the second the entrants. We now have that industry efficiency is less than the harmonic means of the incumbent and entrant efficiencies, $\varepsilon = 1/e = f((x^I + x^E, y^I + y^E), (x, y)) \leq 1/[s^I/f((x^I, y^I), (X, Y)) + s^E/f((x^E, y^E), (X, Y))] = 1/(s^I/\varepsilon^I + s^E/\varepsilon^E)$, where s are the market shares of total incumbent output y^I and total entrant output y^E (both evaluated at the shadow of the industry efficiency program), and x^I and x^E represent total incumbent input and total entrant input. The relative gap, $\varepsilon^E = f((x^I + x^E, y^I + y^E), (X, Y))[s^I/f((x^I, y^I), (X, Y)) + s^E/f((x^E, y^E), (X, Y))]$, measures *entry efficiency*. We now have that *industry productivity growth equals entry efficiency plus mean incumbent and entrant efficiency change plus technical change*:

$$M_t = \varepsilon^E_{t+1} \times \frac{\sum s^i_t/f((x^i_t, y^i_t), (X_t, Y_t))}{\sum s^i_{t+1}/f((x^i_{t+1}, y^i_{t+1}), (X_{t+1}, Y_{t+1}))} \times \sqrt{\frac{f((x_t, y_t), (X_t, Y_t))}{f((x_t, y_t), (X_{t+1}, Y_{t+1}))} \times \frac{f((x_{t+1}, y_{t+1}), (X_t, Y_t))}{f((x_{t+1}, y_{t+1}), (X_{t+1}, Y_{t+1}))}}.$$

It is possible to further disentangle the middle factor, which represents mean incumbent and entrant efficiency change, in individual firm effects and intra-organization effects (of the incumbents and of the entrants).

5.8 Conclusion

In this paper, we have shown that a simple linear program that can be used to measure the efficiency of a firm relative to the industry it belongs to can be applied to the industry as a whole and that the solution decomposes the efficiency of the industry in firm effects and an industrial organization effect. The simple linear programming technique can even be extended to dynamic analysis and then encompasses the measurement of productivity growth and the contributions of a better industrial organization and entry and exit.

Acknowledgments I am grateful to editor Subhash Ray for numerous comments and suggesting to let the static efficiency model encompass productivity growth analysis.

References

Afriat, S.N. 1972. Efficiency estimation of production functions. *International Economic Review* 13(3): 568–598.

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

Baumol, William J. 1977. On the proper cost tests for natural monopoly in a multiproduct industry. *The American Economic Review* 67(5): 809–822.

Baumol, W.J., John C. Panzar, and Robert D. Willig. 1982. *Contestable markets and the theory of industry structure*. New York: Harcourt Brace Jovanovich.

Färe, Rolf, Shawna Grosskopf, and James Logan. 1983. The relative efficiency of illinois electric utilities. *Resources and Energy* 5(4): 349–367.

Førsund, Finn R., and Lennart Hjalmarsson. 1974. On the measurement of productive efficiency. *The Swedish Journal of Economics* 76(2): 141–154.

Francis, Graham, Ian Humphreys, and Jackie Fry. 2005. The Nature and prevalence of the use of performance measurement techniques by airlines. *Journal of Air Transport Management* 11(4): 207–217.

Hicks, J.R. 1935. Annual survey of economic theory: the theory of monopoly. *Econometrica* 3(1): 1–20.

McKenzie, Lionel W. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27(1): 54–71.

ten Raa, T. 2011. Benchmarking and industry performance. *Journal of Productivity Analysis* 36 (3): 285–292.

Sharkey, W.W., and Lester G. Telser. 1978. Supportable cost functions for the multiproduct firm. *Journal of Economic Theory* 18: 23–27.

Yli-Viikari, A., Risku-Norja, H., Nuutinen, V., Heinonen, E., Hietala-Koivu, R., Huusela-Veistola, E., Hyvönen, T., Kantanen, J., Raussi, S., Rikkonen, P., Seppälä, A., and Vehmasto, E. 2002. Agri-environmental and rural development indicators: a proposal. Agrifood Research Reports 5, MAA Agrifood Research Finland.