Chapter 2
Efficiency Measures

2.1 Introduction

In Chap. 1, we introduced efficiency as the use of the fewest inputs (resources) to produce the most outputs (services). This idea is fundamental to much of modern benchmarking literature because it allows us to evaluate performance without clearly defined preferences. That is, we avoid the difficult task of estimating preference functions and deciding on exact priorities. We will expand on this below.

Although the notion of efficiency is simple and intuitive at first glance, there are actually many different ways to conceptualize efficiency. We shall discuss some of the most common concepts in this chapter. We will cover classical concepts from production theory, including technical efficiency, allocative efficiency, and scale efficiency, as well as more advanced concepts like dynamic efficiency and structural efficiency.

Moreover, several of these concepts can be operationalized in different ways. We can, for example, measure technical efficiency in terms of input space, output space, or both types of spaces. We can also measure it in specific directions, etc.

The aim of this chapter is to provide an overview of efficiency-related concepts as well as bits and pieces of the relevant theoretical background.

2.2 Setting

In pursuing this aim, we will generally assume that the technology is given. We focus on a given firm and can therefore describe the setting in the following way: A firm k has used m inputs $x^k = (x_1^l, \dots, x_m^k) \in \mathbb{R}_+^m$ to produce n outputs $y^k = (y_1^k, \dots, y_n^k) \in \mathbb{R}_+^n$. The set of feasible production plans or input-output combinations available to firm k is given by the technology or production possibility set T,

$$T = \{ (x, y) \in \mathbb{R}_+^n \times \mathbb{R}_+^m \mid x \text{ can produce } y \}.$$

P. Bogetoft and L. Otto, *Benchmarking with DEA, SFA, and R*, International Series in Operations Research & Management Science 157, DOI 10.1007/978-1-4419-7961-2_2,

There are many ways to construct the technology T. We have already illustrated some of these methods in Chap. 1, and we will take a closer look at the basic assumptions that one can make about technologies in Chap. 3. Moreover, we shall spend much of the book describing alternative methods like Data Envelopment Analysis DEA and Stochastic Frontier Analysis SFA, which involve the construction of technologies based on actual observations. For now, however, it does not matter how we estimate T. The same efficiency concepts are applicable to technologies estimated in different ways.

2.3 Efficient production

Efficiency is generally a question of using few inputs (resources) to produce many outputs (services).

To be more precise, let us consider two firms, (x^1, y^1) and (x^2, y^2). We say that firm 2 dominates or is more efficient than firm 1 if it uses no more inputs to produce no fewer outputs and is doing strictly better in at least one dimension.

Dominance. (x^2, y^2) dominates (x^1, y^1) if and only if $x^2 \leq x^1$, $y^2 \geq y^1$, and $(x^1, y^1) \neq (x^2, y^2)$

Note that we require the dominating firm, firm 2, to use no more inputs to produce no less outputs than firm 1 and to not be exactly similar to firm 1. Therefore, we require the dominating firm to be strictly better in at least one dimension (to use strictly less of an input or produce strictly more of an output).

Dominance allows us to partially rank firms. Some firms can be compared, while others cannot. This is illustrated in the left panel in Fig. 2.1. firm 2 dominates firm 1, while firm 3 neither dominates nor is dominated by firm 1 or firm 2.

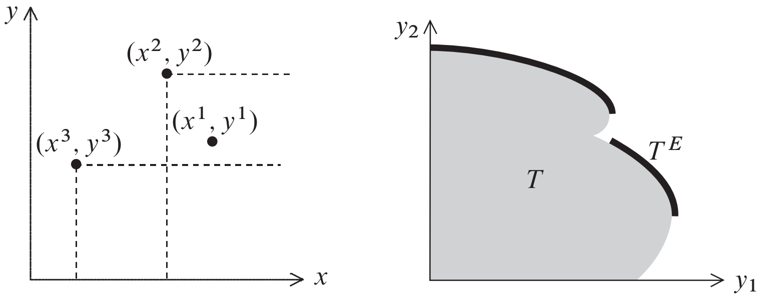


Fig. 2.1 Dominance and efficiency

The dominance relationship is relevant because almost everyone would prefer a more efficient or dominating production plan to the less efficient one that it dominates. For this to hold, we need our preference to be increasing in outputs and decreasing inputs. Thus, dominance is a weak expression of preferences that allows us to partially rank production plans.

In economics, the efficient firms are those that cannot be dominated by other firms. To determine which firms are efficient, we thus need a description of all possible firms (e.g., a listing or a technology set). For a given technology set T, we define efficiency as follows:

Efficiency. (x, y) is efficient in T if and only if it cannot be dominated by some $(x', y') \in T$.

The efficient subset of T, T^E is

$$T^E = \{ (x, y) \in T \mid (x, y) \text{ is efficient in } T \}.$$

The efficient subset T^E of T are the inputs-output combinations that cannot be improved. They represent best practices. An illustration is provided in the right panel of Fig. 2.1. Here the technology set T for 2 outputs is the shaded area, and the efficiency set T^E is the bold part of the frontier. In the production economics literature, this notion of efficiency is sometimes called *Koopmans-efficiency* to distinguish it from other types of efficiency.

The focus on efficiency is natural from a *theoretical perspective*. On the one hand, efficiency is not too strong a requirement; under conditions of mild regularity, one can always identify an optimal production plan from among the efficient ones. On the other hand, we cannot generally strengthen the efficiency requirement; any efficient plan may be the uniquely optimal plan given perfectly sensible underlying but unknown preference functions.

The focus on efficiency is also convenient from an *applied perspective*. One of the main obstacles to the evaluation of effectiveness is to select the objectives or preferences against which we should gauge performance. Here, efficiency provides an easy way out because it only requires that more outputs and fewer inputs are preferable. Thus, instead of engaging in dead-end discussion about overall objectives, we create a partial ranking that will be agreed on by almost everyone. It is worth remembering, however, that this logic also means that while *efficiency is a necessary condition for effectiveness, it is not a sufficient one*. In fact, in terms of a particular technology, an inefficient firm may well be better than a fully efficient one. We could rephrase this by saying that it is not sufficient to run fast; it is also important to run in the correct direction—and it may be better to run at a moderate speed in the right direction than at full speed off-course.

So far, we have defined and explained the relevance of efficiency. We have focused on which firms are efficient and which are not. Additionally, we have introduced a partial ranking of firms in terms of dominance. In the following sections, we will study how to measure efficiency levels. We want to go beyond the efficient/inefficient dichotomy and measure degrees of (in)efficiency.

2.4 Farrell efficiency

The single most widely used approach to measuring the degree of efficiency in a general multi–input and multi–output setting is the strategy suggested by Debreu and Farrell, usually referred to as simply as Farrell efficiency. The idea is to ask if it is possible to reduce the input without changing the output. Seeking to process multiple inputs and outputs in a simple way, we look for a proportional reduction of all inputs.

The *input–based Farrell efficiency* or just *input efficiency* of a plan (x, y) relative to a technology T is defined as

$$E = \min\{\, E > 0 \mid (Ex, y) \in T \,\}$$

i.e., it is the maximal proportional contraction of all inputs x that allows us to produce y. Thus, if $E = 0.8$, it indicates that we could have saved 20% off all inputs and still produced the same outputs.

Likewise, output–based Farrell efficiency or *output efficiency* is defined as

$$F = \max\{\, F > 0 \mid (x, Fy) \in T \,\}$$

i.e., the maximal proportional expansion of all outputs y that is feasible with the given inputs x. Thus, a score of $F = 1.3$ suggests that we could expand the output by 30% without spending additional resources.

A small-scale example of this concept using one input and one output is provided in Fig. 2.2. We see that we can reduce input x to x^* without losing output and that we can increase output y to y^* without using more resources. Therefore, we have

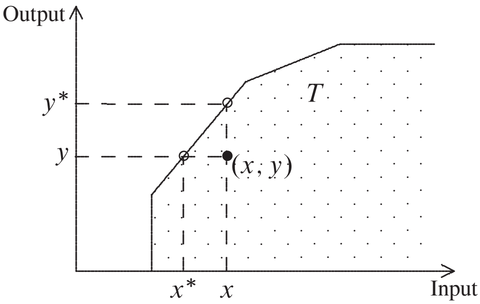


Fig. 2.2 Farrell efficiency in one–input/one–output example

$$E = \frac{x^*}{x},$$
$$F = \frac{y^*}{y}.$$

Table 2.1 Two–input, two–output example

Firm	Input A	Input B	Output C	Output D
1	10	20	20	20
2	20	10	40	20
3	20	30	60	80
4	30	30	80	60
Our	30	20	36	10

Figure 2.3 illustrates how Farrell efficiency is calculated when there are two inputs and two outputs. In the left panel, we show the input isoquant corresponding to the output level y that our firm is producing, and in the right panel, we show the output-isoquant corresponding to the inputs x that our firm is using.

Proportional reduction and expansion correspond to movements along the dashed lines in the two panels. Input efficiency is therefore calculated as the smallest number E that we can multiply on x and remain on or above the isoquant. Likewise, output efficiency is calculated as the largest number F that we can multiply on y and remain below or at the output isoquant. For inputs above and on the input isoquant and outputs below and on the output isoquant curve, we have $E \leq 1$ and $F \geq 1$. The smaller E is and the larger F is, the less efficient the firm is.

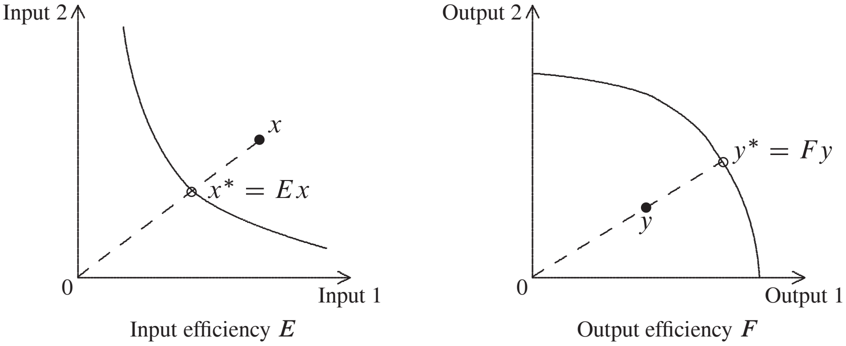


Fig. 2.3 Farrel efficiency

Numerical example

To better understand the logic of the Farrell measures, consider an example in which the technology T is formed using free disposability on the four firms in the upper part of Table 2.1. By this, we mean that any production plan dominated by one of our observed plans is feasible. We are interested in evaluating ”our” firm as given in the last row.

Now we need to look for firms that are using fewer inputs to produce more outputs than our firm. In terms of input usage, we see that only firms 1 and 2 qualify because firms 3 and 4 use too much of input B. At the same time, we see that that firm 1 is not producing enough of output C but that firm 2 produces enough of both outputs. Thus, in effect, firm 2 is really the only firm we can compare with using only dominance (or free disposability).

Now consider the input efficiency of our firm compared to that of firm 2. We see that compared to firm 2, our firm could reduce input A with a factor 20/30 and input B with a factor 10/20. Because it has to decrease in both dimensions, the smallest feasible reduction factor is 20/30. Therefore,

$$E = \frac{20}{30} = 0.67.$$

In a similar way, we see that in terms of output C, our firm could expand with a factor 40/36 by imitating firm 2, and in terms of output D, it could expand with a factor 20/10. Again, because we are looking for the largest possible expansion that works in all output dimensions, we must settle on an expansion of 40/36, i.e.,

$$F = \frac{40}{36} = 1.11.$$

These results are illustrated in Fig. 2.4. In particular, we see that on the input side, it is input A that becomes binding, whereas, on the output side, it is output C that becomes the limiting product.

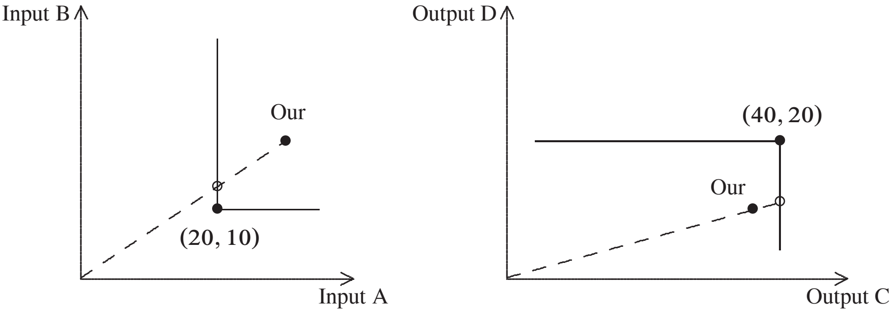


Fig. 2.4 Illustration of numerical example

2.4.1 Non-discretionary inputs and outputs

In applications, we often have situations in which some of the inputs or some of the outputs are fixed and uncontrollable, at least in the short run or using the discretionary power of the firm or unit that we seek to evaluate. A very simple but useful way to handle such situations is to only look for improvements in the discretionary (controllable) dimensions. In this way, if we divide the input and outputs into variable (v) and fixed (f) inputs and outputs as in $(x, y) = (x_v, x_f, y_v, y_f)$, we can define the input and output variants of the Farrell measures as follows:

$$E^* = \min\{\, E > 0 \mid (Ex_v, x_f, y_v, y_f) \in T \,\}$$
$$F^* = \max\{\, F > 0 \mid (x_v, x_f, Fy_v, y_f) \in T \,\}.$$

E^* indicates that we can proportionally reduce all variable inputs x_v with a factor E^* without using more of the fixed inputs x_f and without producing fewer of the outputs $y = (y_v, y_f)$. Likewise, the interpretation of F^* is that we can proportionally expand all variable outputs y_v without reducing any of the fixed outputs y_f and without using more inputs than $x = (x_v, x_f)$.

2.4.2 Using Farrell to rank firms

The main outcome of many benchmarking studies is a list of the Farrell efficiency values of the firms in an industry. Such lists or league tables are studied with interest and care because they are often considered to provide firm rankings with the best firms having the largest E and the worst having the lowest E (or vise versa for F).

One can discuss whether efficiency measures can really be used to rank firms or whether they solely provide individual measures of efficiency and thus improvement potential.

Purists would argue that rankings using Farrell efficiency are only possible to a very limited degree. A case can be made only for comparing firms where one dominates the other. In such situations, the efficiency score achieved by comparing one to the other is simply a way to quantify the amount of dominance.

One can also take a more pragmatic view and argue that even in cases in which the two units are not comparable based on dominance, the Farrell measure still provides a performance measure, and that low Farrell efficiency is an indication of high improvement potential. It is important to note, however, that this is not a simple ranking in which everyone is competing with everyone. It is more similar to a golf tournament or a horse race with handicapping. The technology defines the performance standard for each of the firms, and only hereby can we compare firms that produce different service mixes or use different input mixes.

Still, this use of efficiency scores presumes that the inputs and outputs correctly characterize the available options and that we do not have any more information about the relative importance of the different inputs and outputs.

Additionally, in reality, the technology may be described more precisely in some parts of the production space than others, and we shall talk at some length about bias in DEA in Chap. 5. Such biases in our descriptions make it less reasonable to consider the ranking as the result of a fair race. Indeed, in such cases, one can argue that not all firms participate in the same race and that the rankings are not fair because it is easier to be close to the winner in some races (with fewer or less talented competitors) than in other races. These differences make it difficult to make comparisons across races. (The second-best runner in Denmark is probably quite a bit worse than the second-best US runner).

It is also worthwhile to note that Farrell efficiency is not exactly the same as traditional (Koopmans) efficiency as introduced in Sect. 2.3. That is, $E(x, y; T) = F(x, y; T) = 1$ does not imply $(x, y) \in T^E$. This situation occurs when some inputs can be reduced and/or some outputs can be expanded individually but there is no option to contract or expand all inputs or outputs simultaneously (i.e., when we are on a horizontal or vertical part of the isoquants). This is one drawback of the Farrell measure.

2.4.3 Farrell and Shephard distance functions

Farrell efficiency depends on our starting point (x, y) and the technology set T. Instead of using E and F above, it would thus be more precise to use the longer notation $E((x, y); T)$ and $F((x, y); T)$. In many contexts, however, this would be too cumbersome and we simply use E and F or perhaps $E(x, y)$ and $F(x, y)$. In some cases, we also call these efficiency measures *distance functions* or, more precisely, input distance functions and output distance functions. This nomenclature emphasizes that they are not just numbers but are also procedures (functions) that map technologies and observations into real numbers.

Some prefer to work with the so-called Shephard measures rather than the Farrell measures. For the sake of completeness, we note that the *Shephard distance functions* are simply the inverse of the Farrell ones,

$$D_i(x, y) = \max\{\, D > 0 \mid \big(\frac{x}{D}, y\big) \in T \,\} = \frac{1}{E(x, y)}$$

$$D_o(x, y) = \min\{\, D > 0 \mid \big(x, \frac{y}{D}\big) \in T \,\} = \frac{1}{F(x, y)}.$$

The function D_i is called the (Shephard) input distance function and D_o the (Shephard) output distance function. Some computer programs calculate these functions rather than the Farrell variants.

2.5 Directional efficiency measures

In the Farrell (and Shephard) approach to efficiency measurement, all inputs are reduced or all outputs are expanded by the same factor. This proportional adjustment has been challenged by a series of alternative efficiency measurements approaches. We cover a few of these here.

An early suggestion in the DEA literature was to consider simultaneous improvements on the input and output side by basically combining the Farrell input and output efficiency measures into one measure, sometimes referred to as the *graph hyperbolic measure of technical efficiency*

$$G = \min\{\, G > 0 \mid (Gx, \frac{1}{G}y) \in T \,\}.$$

In G, we seek to simultaneously reduce inputs and expand outputs as in the Farrell approach. The input side is exactly as in the E measure, and the output side is in the spirit of the F measure; when we reduce G, we expand $1/G$, which is like the F factor is in the Farrell output efficiency measures. Also note that for $(x, y) \in T$, we have $G \leq 1$.

The interpretation of a graph hyperbolic efficiency G is that we can make due with input Gx and simultaneously expand output to $\frac{1}{G}y$. This is illustrated in Fig. 2.5. The curve traversed by $(Gx, \frac{1}{G}y)$ when G takes all positive values is a hyperbola; this is indicated by the dashed line, and by comparing the intersection and the original point, we can measure G on either the input or the output axis as indicated in the figure.

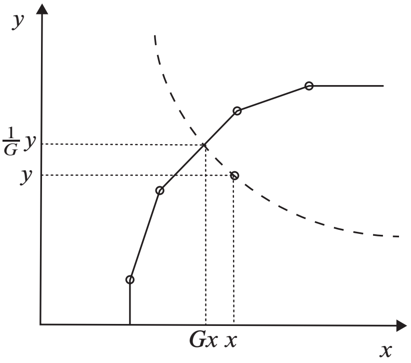


Fig. 2.5 Graph efficiency measure

In applications, G is not always easy to implement because of the non-linearities involved.

A more profound alternative or generalization of Farrell's proportional approach is based on *directional distance functions*. We will discuss this approach now, and to simplify the exposition, we initially focus on the input side.

The purpose of directional distance functions is to determine improvements in a given direction $d \in \mathbb{R}^m_+$ and to measure the distance to the frontier in such d-units. This process leads to a directional distance or *excess function*

$$e = e(x, y; T, d) := \max\{ e \in \mathbb{R}_+ \mid (x - ed, y) \in T \}.$$

The excess $e(x, y; T, d)$ has a straightforward interpretation as the number of times the input bundle d has been used in x in excess of what is necessary to produce y. Therefore, a large excess reflects a large (absolute) slack and a considerable amount of inefficiency. It shows how many times we can harvest the improvement bundle d if we were to learn best practice.

An illustration is provided in Fig. 2.6 for 2 different directions $(1, 0.25)$ and $(.25, 4)$ in addition to the usual Farrell direction. In the figure, we have also indicated the projection points using circles. On this basis, the efficiency figures can be calculated.

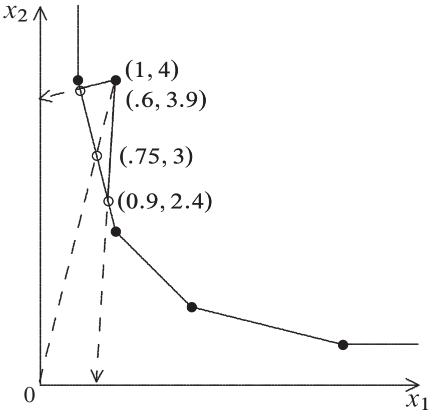


Fig. 2.6 Input directional efficiency

The directional distances and the Farrell input efficiency in this case become

$$e((1,4); (1, 0.25)) = 0.48$$
$$e((1,4); (0.25, 4)) = 0.39$$
$$E((1,4)) = 0.75$$

We note that the directional distances are not comparable across different directions. Excess values depend on the directions in which we move, as this example shows. This dependence means that we need to exercise care in interpreting the results and particularly in comparing excess values across firms and directions. On the other hand, it is also a useful property because by measuring the excess in different directions, one can get a picture of which particular resources a firm seems to have in excess. This approach is sometimes called multidirectional efficiency analysis, MEA, and can be used to select improvement strategies, etc.

The excess values also depend on the length of the direction vector. Thus, for example, if we double the length of the improvement direction d, the number of times we can save the doubled vector is halved. More generally, for arbitrary $\theta \geq 0$, we have

$$e((x, y); \theta d) = \frac{1}{\theta} e((x, y); d).$$

Again, this simply requires us to be explicit in interpreting the results and making comparisons across different firms and in different directions.

The Farrell approach is, in principle, just a special variant of the directional distance function approach, where we use the firms own inputs as the direction vector. Thus, it is straightforward to see that

$$e((x, y); x) = 1 - E(x, y)$$

That is, with direction equal to what is present in the existing input production plan, the excess function is a measure of the inefficiency of the firm as determined using the Farrell method. If Farrell efficiency is 80%, for example, the excess is 20%. Likewise, the Farrell efficiency measures when some of the inputs or outputs are fixed are special variants of the directional distance approach. We have, for example,

$$e((x, y); (x_v, 0)) = 1 - E^*(x_v, x_f, y).$$

Rather than creating a direct dichotomy between controllable and non-controllable elements, the directional distance function approach allows us to work with grades of discretion—some dimensions can be controlled more easily than others, and some dimensions are more desirable to change than others.

Like in the graph efficiency, we can combine the input and output efficiency perspectives using the direction distance function approach. That is, we can examine whether it is possible to use fewer inputs and produce more outputs. Thus, we can look for changes in the direction $(d_x, d_y) \in \mathbb{R}^m_+ \times \mathbb{R}^n_+$ and define the directional excess e as

$$e = \max\{ e > 0 \mid (x - ed_x, y + ed_y) \in T \}.$$

With one input and one output and the direction $(d_x, d_y) = (1, 1)$, we have the situation indicated in Fig. 2.7, where the arrow indicates the direction.

An important question, in theory as well as in practice, is which direction is best. The correct (but also somewhat easy) answer at this stage is that it depends on the application. We will return to this question in Sect. 2.9 below because it is also related to the question of using input– or output–based efficiency measures, technical or allocative efficiency measures, etc.

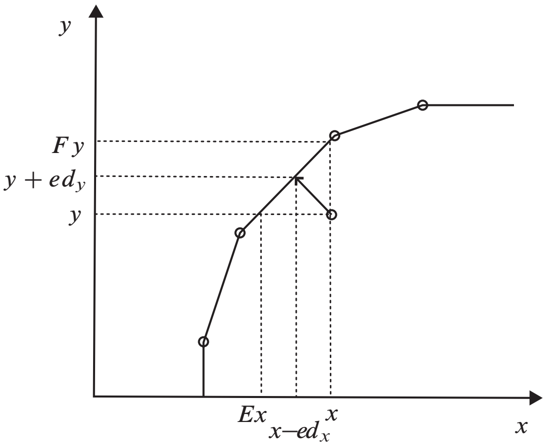


Fig. 2.7 Input and output directional efficiency

Practical application: Benchmarking in waterworks

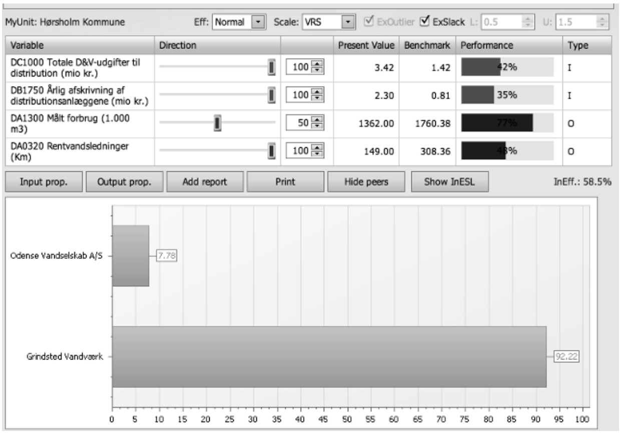
Rather than theoretically discussing the pros and cons of different directions, one can take a pragmatic approach and see the direction as a steering instrument by which a user can control the projection of a firm on the efficient frontier. This approach in used in the interactive benchmarking system called IBEN that is used by Danish Waterworks, cf. also Sect.1.1.1, to support individual learning among waterworks. An illustration is provided in Fig. 2.8. In the specific model (technology), it is presumed that the evaluated waterworks have used two inputs to produce two outputs. The inputs are the Opex (DC1000) and Capex (DB1750) measures, and the outputs are the water distributed (DA1300) and length of the water lines (DA0320). We see, for example, that Hørsholm has used 3.43 million DKK of Opex and 2.3 mio DKK of Capex to distribute 1.362 million m^3 of water and maintain 149 km of waterlines. 1 million DKK is approximately 150 thousand Euro.

In the illustration, we see that the user has chosen to look for improvements in all directions (i.e., improvements to both inputs and outputs). The first output, however, is emphasized less than the other outputs. The sliders work to choose the direction and thereby steer the projection of the analyzed firm onto the efficient frontier. The figures that indicate direction in IBEN are percentages, and the idea is that they are percentages of the present values for the analyzed firm. Therefore, the correspondence between the IBEN illustration and our framework here is as follows:

$$d = (100\%3.42, 100\%2.30, 50\%1362, 100\%149) = (3.42, 2.3, 681, 149)$$

The resulting benchmark is also shown in IBEN. In our notation, the natural benchmark would be

$$\text{Benchmark} = (x - ed_x, y + ed_y)$$



MyUnit: Hørsholm Kommune   Eff: Normal   Scale: VRS   ☑ ExOutlier   ☑ ExSlack   L: 0.5   U: 1.5

Variable	Direction		Present Value	Benchmark	Performance	Type
DC1000 Totale D&V-udgifter til distribution (mio kr.)		100	3.42	1.42	42%	I
DB1750 Årlig afskrivning af distributionsanlæggene (mio kr.)		100	2.30	0.81	35%	I
DA1300 Målt forbrug (1.000 m3)		50	1362.00	1760.38	77%	O
DA0320 Rentvandsledninger (Km)		100	149.00	308.36	48%	O

InEff.: 58.5%

Fig. 2.8 IBEN evaluation of Danish waterwork

but this is not exactly the same as what is shown in the illustration, except with regard to Opex. The reason for this difference is that this benchmark may contain slack and that the slack has been eliminated from the example, cf. also the ExSlack checkbox.

Lastly, we note that IBEN shows the excess value e as InEff and the individual inefficiencies on the different dimensions. Thus, for example, a value of 35% on the Capex line shows that it is only necessary to use 35% of the present Capex level (i.e., 0.81/2.30=35%). In addition, IBEN provides information about the peer units on the lower part of the screen. In this way, the user can see which entities to learn from and how this depends on the direction chosen. IBEN also allows the user to easily remove peers and hereby to re-estimate the technology and directional efficiency on a modified technology.

2.6 Efficiency measures with prices

So far, we have focused on performance evaluations in contexts with a minimum of information. We have assumed that we have firms transforming multiple inputs $x \in \mathbb{R}^m_+$ into multiple outputs $y \in \mathbb{R}^n_+$ using one of several possible production plans T. In addition, we have assumed that we prefer more outputs and fewer inputs. Except for this assumption, we have made no assumptions about the relative importance of inputs and outputs.

In some situations, however, we know a priori the relative weights, prices or priorities that can naturally be assigned to the different inputs and/or outputs. Such information allows us to make more focused evaluations. Moreover, it allows us to decompose efficiency into technical efficiency, associated with the use of optimal

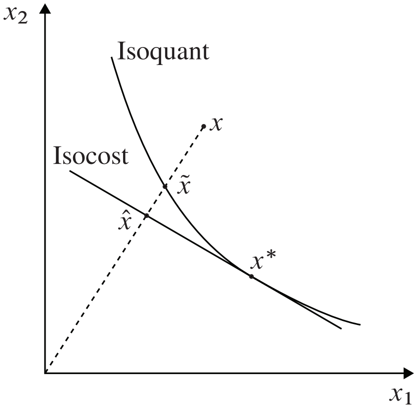


Fig. 2.9 Cost minimum

procedures, and allocative efficiency, associated with the use of optimal combinations of inputs or the production of optimal combination of outputs.

2.6.1 Cost and input allocative efficiency

Let us consider an example in which we have prices associated with the inputs. Let w be the n–the vector of input prices, $w \in \mathbb{R}^n_+$.

In this situation, we can calculate the costs wx of a given production plan (x, y), and thereby evaluate the production plan (x, y) via the cost output combination (c, y), where $c = wx$. In principle, we can conduct efficiency analyses of this, more aggregated, model just as we did with the (x, y) model.

It is now intuitively clear that it is easier to be efficient using the (x, y) model than the (c, y) model because in the latter situation, the firm is responsible not only for picking a technically efficient point on T^E but also for picking the right one to minimize the costs. We shall refer to the latter as the allocation problem and the associated efficiency as *allocative efficiency*.

To formalize this idea, let us assume that a firm has used inputs x, as illustrated in Fig. 2.9.

Ignoring the price information, we can measure Farrell efficiency in the usual way. To distinguish it from other forms of efficiency here, we will now call this the *technical input efficiency* of observation x. As we have seen, it is the maximal contraction of the input bundle and can be calculated as

$$TE = \frac{\|\tilde{x}\|}{\|x\|}$$

where $\tilde{x}$ is the point on the isoquant obtained via proportional scaling for the observed x along the dashed line in the figure.

In the same way, we can measure *cost–efficiency* CE as the ratio between the minimal cost and the actual cost

$$CE = \frac{wx^*}{wx}.$$

The optimal minimal cost input combination x^* is found by solving the cost minimization problem

$$\min\ w'x \quad \text{subject to} \quad (x', y) \in T.$$

The solution to this optimization problem is well known to be the point x^* where the isocost line is tangent to the isoquant as shown in Fig. 2.9.

Cost–efficiency CE is actually also Farrell efficiency in the more aggregate model that uses costs as inputs.

Before we proceed, let us rewrite technical efficiency, TE. It is clear that technical efficiency is also equal to the cost of $\tilde{x}$ compared to the cost of x because the two vectors are proportional. That is, because $\tilde{x} = TEx$, we also have $w\tilde{x} = TEwx$, and therefore

$$TE = \frac{w\tilde{x}}{wx}$$

If we can save 20% of all inputs from x to $\tilde{x}$, we can also save 20% in costs.

Now compare the costs of $\tilde{x}$ and x^*. The difference is the cost of having picked the technically efficient plan $\tilde{x}$ rather than another and less expensive input mix x^*. Thus, the difference represents an allocation problem, and we define *allocative efficiency* as

$$AE = \frac{wx^*}{w\tilde{x}}$$

We see that $AE \leq 1$. If, for example, AE is 0.8, it means that we could have saved 20% by better allocating our funds toward a less expensive but sufficient input mix.

In summary, we now have three different efficiency measures: technical efficiency TE, cost efficiency CE and allocative efficiency AE. The relationship between them is easy to derive:

$$CE = \frac{wx^*}{wx} = \frac{wx^*}{w\tilde{x}}\,\frac{w\tilde{x}}{wx} = AE \cdot TE$$

This decomposition emphasizes our initial intuition. *To be cost–efficient, the firm must be able to choose the right mix of inputs and use them in a technically efficient manner*. It must use the right resources, and it must use them in the right way.

In closing, we note that if we define $\hat{x}$ as the point on the dotted contraction curve that has the same costs at x^*, then we can also look at the relationship between CE, AE and TE by comparing the length of x, $\tilde{x}$ and $\hat{x}$ as follows:

$$TE = \frac{\|\tilde{x}\|}{\|x\|}, AE = \frac{\|\hat{x}\|}{\|\tilde{x}\|}, CE = \frac{\|\hat{x}\|}{\|x\|}$$

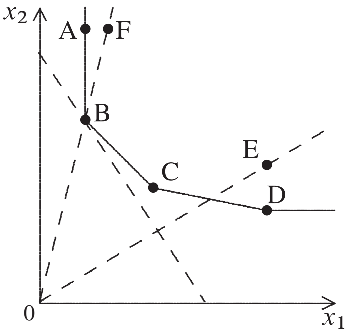


Fig. 2.10 Technology for cost minimization example

i.e., by comparing the lengths of vectors on the dotted line. We see that all of these efficiency measures are smaller than or equal to 1.

Numerical example

Consider a simple example in which six firms A – F have used two inputs to produce one output. The data are provided in Table 2.2.

Table 2.2 Data for cost minimization

Firm	x_1	x_2	y	Costs wx
A	2	12	1	15.0
B	2	8	1	11.0
C	5	5	1	12.5
D	10	4	1	19.0
E	10	6	1	21.0
F	3	12	1	16.5
Price w	1.5	1.0		

We see that all firms have produced the same output, so we can safely look at the problem in the input space. Assuming free disposability and convexity (i.e., that we can produce at least the same outputs with more inputs and that weighted averages (line segments between observations) are feasible as well) we can construct a technology from these observations. The input isoquant (for $y = 1$) of this technology is illustrated in Fig. 2.10 below. The assumptions of free disposability and convexity will be discussed in detail in the next chapter.

The resulting efficiency values are shown in Table 2.3. We see that all firms except E and F are on the frontier and thus are technically efficient; i.e., they have $TE = 1$. The technical efficiency of firms E and F can be calculated by first noting

Table 2.3 Economic efficiency

Firm	*CE*	*TE*	*AE*
A	0.73	1.00	0.73
B	1.00	1.00	1.00
C	0.88	1.00	0.88
D	0.58	1.00	0.58
E	0.52	0.75	0.70
F	0.67	0.67	1.00

that they are projected onto $0.5C + 0.5D = (7.5, 4.5)$ and $B = (2, 8)$ respectively. Thus, for example, the *TE* of F is $2/3 = 8/12 = 0.66$.

Although most of the firms are technically efficient, they have not been equally good at selecting the cost-minimal input mix. These differences become clear when we calculate costs in the last column of Table 2.2. We see that the firm with the lowest costs is B, with a cost value of 11. This result is not surprising given Fig. 2.10, in which the isocost curve is tangent to the isoquant at B. Calculating cost efficiency is now also straightforward. Thus, for example, the cost efficiency for firm A is $CE = 11/15$ because cost efficiency is the minimal cost compared to the actual cost. It is similar to the technical efficiency measure except that we make the evaluation using a one-input (costs) framework.

Lastly, having calculated both *TE* and *CE*, we can easily determine allocative efficiency, defined as $AE = CE/TE$.

We also note that F is allocatively efficient but not technically efficient. This is the case because F is projected onto the cost minimal production plan B when we remove technical efficiency. The classical approach to allocative efficiency that we have introduced here requires one always to measure allocative efficiency at the frontier.

2.6.2 Revenue and output allocative efficiency

A parallel treatment of allocative issues is possible on the output side. Here we look at whether the output mix is optimal in terms of maximizing revenue for a given input. This depends on the output prices $p \in \mathbb{R}^m_+$. An illustration is provided in Fig. 2.11.

As with cost efficiency, we can define revenue efficiency as

$$RE = \frac{py^*}{py}$$

where y is the observed output and y^* the optimal revenue output i.e., the solution to the revenue-optimizing problem

$$\max \; py' \quad \text{subject to} \quad (x, y') \in T.$$

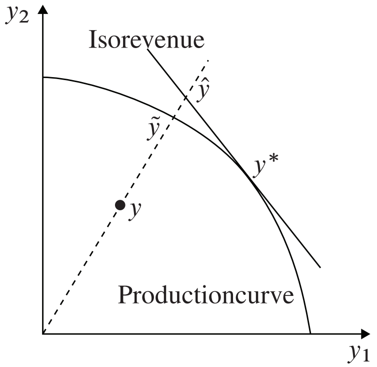


Fig. 2.11 Revenue maximum

We can now rewrite revenue efficiency as

$$RE = \frac{py^*}{py} = \frac{py^*}{p\tilde{y}} \frac{p\tilde{y}}{py} = \frac{py^*}{p\tilde{y}} F = AF \cdot F.$$

Here $\tilde{y} = Fy$, and F is the Farrell output technical efficiency. Therefore, $\tilde{y}$ is the technically efficient point that we obtain when we expand y radially along the dotted line. Also note that we have introduced a shorthand AF for output-oriented allocative efficiency:

$$AF = \frac{py^*}{p\tilde{y}}.$$

Output allocative efficiency is the revenue obtained by choosing the best mix of output relative to the revenue from simply being technically efficient.

To be fully revenue-efficient, a firm must demonstrate both full output technical efficiency and full output allocative efficiency. It must use the best procedures to get the most out of its resources, and it must produce the right mix of services. This concept is sometimes summarized by saying that *it is not enough to do things right; one must also do the right things.*

As in the analyses of the input side, we can also look at this decomposition of revenue efficiency in terms of vector lengths. To see this, let us define $\hat{y}$ as the point on the dotted expansion line that has the same revenue as y^*. We then have

$$F = \frac{\|\tilde{y}\|}{\|y\|}, AF = \frac{\|\hat{y}\|}{\|\tilde{y}\|}, RE = \frac{\|\hat{y}\|}{\|y\|}$$

i.e., by comparing the lengths of vectors on the dotted line, we can calculate all three efficiency values.

2.6.3 Profit efficiency

If we have prices w and p on both the input and the output side, we can, of course, also evaluate the firms ability to generate profit and use this as the benchmarking focus. In such situations, we will naturally define *profit efficiency* as

$$PE = \frac{py - wx}{py^* - wx^*}$$

where (x, y) is the observed production plan and (x^*, y^*) is the profit- maximizing production plan, i.e., the solution to

$$\max \; py' - wx' \quad \text{subject to} \quad (x', y') \in T.$$

A small value of PE would be an indication that large profit potentials have been foregone.

Again, one can decompose the inefficiency into different parts related to 1) technical inefficiency, 2) input allocative efficiency and 3) output allocative efficiency. All of these different forms of efficiency describe the firms ability to get the most out of given resources, select a cost-minimal input mix, and select a revenue-maximizing output mix. The decomposition will be somewhat arbitrary depending on the order in which we identify the elements and particularly on the choice of an input- or output-oriented technical efficiency measure. We will not discuss the alternatives in any more detail here.

2.7 Dynamic efficiency

Over time, the behavior and performance of firms are likely to change. We need measures that capture such changes. In addition, the technology is likely to change due to technical progress. These changes make it relevant to measure not only how firms change over time but also how many of these changes are caused by general technological progress and how many can be attributed to special initiatives on the part of individual firms that improve relative to the existing technology.

An example of these dynamic issues is provided in Fig. 2.12 below. We depict the state of one firm during two periods: first, period s and then period t. Likewise, we have two technologies that are relevant for the two periods. We see that the firm has improved in the sense that from s to t, it has moved closer to the s technology. On the other hand, the technology has also shifted, which has made it less costly to produce. Therefore, the firm has not improved as much as we would expect from a general technological development perspective. In period t, it has more excess costs than in period s.

In the benchmarking literature, the most popular approach to dynamic evaluations is the Malmquist index. It works without prices to aggregate the different inputs and outputs.

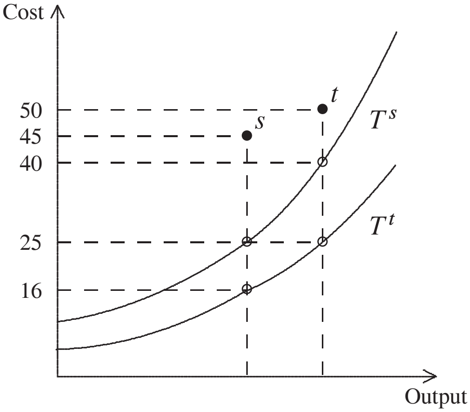


Fig. 2.12 Dynamic change of performance and technology

To explain the idea of the Malmquist index, let $E^i(s,t)$ be a *measure of the performance of firm i in period s against the technology in period t*. It might, for example, be Farrell input efficiency, i.e., $E^i(s,t) = \min\{ E > 0 \mid (Ex^{is}, y^{is}) \in T^t \}$. However, it could also be other measures, including output-based ones, as long as larger values reflect better performance (closeness to the frontier). Note that we now distinguish the technology and the production data depending on the period from which they stem. In this paragraph, we will focus on the performance of firm i, and therefore, we can simplify the notation and use $E(s,t)$ instead of $E^i(s,t)$.

To measure the improvement in firm i from period s to period t, we can look at the changes in efficiency compared to a fixed technology. If we use time s technology as our benchmark, we can look at

$$M^s = \frac{E(t,s)}{E(s,s)}.$$

If the firm has improved from period s to t, $E(t,s) \geq E(s,s)$, and therefore, M^s is larger than 1. If, on the other hand, the firm is less efficient in period t than period s, $E(t,s) \leq E(s,s)$, the ratio is smaller than 1. Therefore, M^s is larger than 1 when the firm improves over time and smaller than 1 if it moves away from the frontier over time. For example, if a firm is 40% efficient in period s and 80% efficient in period t, then it has improved by a factor of 2. This is a natural way to measure the change. If the firm is always producing the same outputs and if it is using the same input mix in the two periods, then in the example, it must have halved its use of resources in period t compared to period s to show this kind of change in its efficiency score. Of course, real problems are more complicated because the input mix and output mix are likely to change as well, but the interpretation of the ratio is still basically the same.

M^s measure the improvement relative to technology s. We might alternatively have used technology at time t as the fixed technology, in which case we would then get

$$M^t = \frac{E(t,t)}{E(s,t)}$$

Because there is no reason to prefer one to the other, the Malmquist index is simply the geometric average of the two:

$$M(s,t) = \sqrt{M^s\, M^t} = \sqrt{\frac{E(t,s)}{E(s,s)}\,\frac{E(t,t)}{E(s,t)}}.$$

The Malmquist index measures how much a firm has improved from one period s to the next t. The change in performance may, however, be due to two reinforcing or counteracting factors: the general technological progress (or regression) that we would expect everyone to be affected by and special initiatives in the firm that have enabled it to change its performance relative to that of the other firms. We can decompose the Malmquist measure in these two effects by rewriting M as follows:

$$M(s,t) = \sqrt{\frac{E(t,s)}{E(t,t)}\,\frac{E(s,s)}{E(s,t)}}\,\frac{E(t,t)}{E(s,s)} = TC(s,t)\,EC(s,t)$$

where

$$\begin{aligned} TC(s,t) &= \text{ technical change } = \sqrt{\frac{E(t,s)}{E(t,t)}\,\frac{E(s,s)}{E(s,t)}} \\ EC(s,t) &= \text{ efficiency change } = \frac{E(t,t)}{E(s,s)} \end{aligned}$$

The *technical change index, TC* is the geometric mean of two ratios. In both, we fix the firm's production plan at time t and use this as the fixed point against which we measure changes in the technology. If the technology has progressed, we will have $E(t,s) > E(t,t)$ because the technology has moved further away from the given observation (i.e., the first ratio in the geometric mean is > 1). The idea of the second ratio is the same; here we just use the time s version of our firm as the fixed point when we look at technological developments. In summary, the TC measures technological change, and values above 1 represent technological progress in the sense that more can be produced using fewer resources.

The other factor is the *efficiency change index EC*, which measures the *catch-up* relative to the present technology. We always measure this factor against the present technology, asking if the firm has moved closer to the frontier. If so, $E(t,t) > E(s,s)$, and the ratio is larger than 1.

The Malmquist measure is useful to us in understanding how benchmarking results change over time. A firm that has made improvements over the course of a year may be frustrated to learn that it is actually coming out worse in a new benchmarking analysis. The point is, however, that it is not sufficient for a firm to improve compared to itself. The firm must also improve relative to others, and they have also benefited from general technological progress. Thus, the only way to improve is to catch up to the best, i.e., to get closer to the frontier.

The Malmquist measure and its decomposition are useful in capturing dynamic developments from one period to the next. One should be careful in interpreting results from several periods. One cannot simply accumulate the changes because the index does not satisfy what is called the *circular test*; i.e., we may not have $M(1,2)\cdot M(2,3) = M(1,3)$ unless the technical change is particularly well-behaved (Hicks-neutral). This drawback is shared by many other indices and can be remedied by, for example, using a fixed-base technology.

Lastly, let us mention that some of the ideas in the Malmquist approach can also be used to determine the effects of other changes besides time. We could, for example, let s and t represent two different ways to organize production, two different countries, or two technologies, one with and one without advanced automation (robots). The technological change (TC) in such situations would then reflect the general impact of the technological opportunities created by using alternative organizational forms, operating in one or another country or introducing the use of robots.

Numerical example

To give an example of how the formula is used, let us calculate M, TC and EC in the example shown in Fig. 2.12. Using Farrell input efficiency and observing that the inputs in the example are shown on the vertical axis, we can observe the following directly from the graphs:

$$\underset{\text{(Malmquist index)}}{M(s,t)} = \sqrt{\frac{E(t,s)}{E(s,s)}\,\frac{E(t,t)}{E(s,t)}} = \sqrt{\frac{40/50}{25/45}\,\frac{25/50}{16/45}} = \sqrt{\frac{81}{40}} = 1.423$$

$$\underset{\text{(Technical change)}}{TC(s,t)} = \sqrt{\frac{E(t,s)}{E(t,t)}\,\frac{E(s,s)}{E(s,t)}} = \sqrt{\frac{40/50}{25/50}\,\frac{25/45}{16/45}} = \sqrt{\frac{5}{2}} = 1.581$$

$$\underset{\text{(Efficiency change)}}{EC(s,t)} = \frac{E(t,t)}{E(s,s)} = \frac{25/50}{25/45} = 0.9$$

This illustrates what can also be inferred from the graph: the firm has improved from period s to t. If we fix the technology, we see that it has moved much closer to the minimal cost curve. The Malmquist index suggests a 42.3% improvement. What is also clear, however, is that this improvement should be expected simply on the basis of the technological developments. In fact, the frontier shift generates a 58.1% improvement cost. So, the firm has not quite been able to follow the trend of technological development but has instead fallen back an additional 10%. The EC and TC effects are multiplicative, such that $EC \cdot TC = M$.

Practical application: Regulation of electricity networks

Most European distribution companies, DSOs, are regulated by competition authorities, cf. Sect. 1.1.3. The single most widely used type of regulation is the revenue-cap regulation, in which the regulator defines ex ante the maximal allowed price companies can change their consumers over the next 3-5 years. A typical scheme would be

$$R^i(t) = C^i(0)Q(0,t)P(0,t)(1-x-x^i)^t, \qquad t = 1,\dots,5$$

where $R^i(t)$ is the allowed revenue in period t for firm i, $C^i(0)$ is the actual cost of running the DSO in period 0, $Q(0,t)$ is a quantity index reflecting the increase in services from time 0 to t, $P(0,t)$ is a similar index reflecting changes in prices (inflation), and x is a general requirement imposed on all firms and x^i is a specific, additional revenue reduction requirement imposed on DSO i. Hence, the idea is that the regulator allows the DSO to cover its costs but, on a yearly basis, requires it to conduct a general cost reduction of x (e.g., 1.5%) and a specific cost reduction of x^i (e.g., 3%). The advantage of this scheme is that it allows firms to keep what they gain by cutting costs (at least beyond the $x + x^i$ requirement), thus providing them with proper incentives. Also the scheme protects consumers against excessive changes by ex ante requiring charges to fall (with $x + x^i$).

In the implementation of these schemes, a major issue is now how to determine general and individual requirements, x and x^i, respectively. In most cases, solving this problem requires the use of advanced benchmarking. Indeed, x is often established as the frontier shift in Malmquist analyses run on data from a period of some 3-5 years prior to the regulation. Thus, if $TC = 1.02$, the regulator will set $x = 2\%$. Likewise, the setting of x^i is typically informed by a benchmarking model covering, for example, the period $t = 0$ or $t = -1$. The typical benchmarking study will calculate the cost efficiency of each firm and then decide how many years the firm should have to eliminate its incumbent inefficiency, i.e., how quickly it must catch up to best practice. Thus, for example, if a firm has cost efficiency of $CE = 0.80$, it might be asked to partially eliminate this advantage during the regulation period via an extra yearly reduction in costs of, for example, $x^i = 3\%$.

Similar schemes are used to regulate many other sectors as well as to guide budget allocation in public and private organizations.

2.8 Structural and network efficiency

Most of the benchmarking literature is concerned with evaluating the performance of individual firms, i.e., the unit of analysis is firms. It is, however, also possible to evaluate the efficiency of a collection of firms and thus to evaluate if we have the best possible industry structure or if it would pay to move production around, perhaps merging some of the firms and splitting up others. We will briefly illustrate how such analyses can be conducted and return to more comprehensive and complicated cases in later chapters.

First, consider the possible impact of merging firms 1 and 2, which have used similar inputs to produce similar outputs (i.e., a horizontal merger). Let their present production be (x^1, y^1) and (x^2, y^2), respectively. We do not require that they use exactly the same input and output types because we can always allow the value of some of the dimensions of the x and y vectors to be 0.

If the two units become integrated but continue to operate as two independent entities, they will transform the vector of inputs $x^1 + x^2$ into the vector of outputs $y^1 + y^2$. To evaluate the potential efficiency gains from the merger, we can therefore evaluate the efficiency of the latter transformation, i.e., the use of $x^1 + x^2$ to produce $y^1 + y^2$.

Using a Farrell input approach provides us with the following measure of the potential gains from merging firms 1 and 2:

$$E^{1+2} = \min\{E \in \mathbb{R}_+ \mid (E(x^1 + x^2), y^1 + y^2) \in T\}\,.$$

Here E^{1+2} is the maximal proportional reduction in the aggregated inputs $x^1 + x^2$ that allows the production of the aggregated output $y^1 + y^2$.

If $E^{1+2} < 1$, we can save via a merger. If $E^{1+2} > 1$, the merger is costly. A score of $E^{1+2} = 0.8$ would suggest that 20% of all inputs could be saved by integrating firms 1 and 2. Likewise, a score of $E^{1+2} = 1.3$ would suggest that integration would necessitate 30% more of all resources. We shall investigate such measures and conduct some useful decompositions in more detail in Chap. 9.

Practical application: Merger control in health care

The evaluation of potential gains from mergers is used in Dutch regulations to shape the health authorities view of proposed mergers. If two hospitals merge, the competition in the sector decreases, and this will generally decrease the quality of care. Industrial economics models of imperfect competition are used to quantify the likely negative market effects. On the other hand, a merger may also be sufficiently efficiency-enhancing and cost-reducing to be attractive despite the reduced competition. To quantify the possible efficiency gains, the Dutch health authority has estimated models of hospital production and set up evaluations of gains like E^{1+2} above. If E^{1+2} is sufficiently small, this will sway the evaluators in favor of allowing the merger.

Rather than merging two or more firms, which may be costly—especially if the technology shows decreasing returns to scale—we can also try to preserve the existing number of firms and simply reallocate production between them. The potential gains from this step can be calculated in the following way. Imagine that we have 3 firms. Generalizations to more firms are straightforward. Let the firms be denoted $k = 1, 2, 3$, and let their original productions be (x^k, y^k), $k = 1, 2, 3$. Assume that we pick new production plans (x^{*k}, y^{*k}) for each $k = 1, 2, 3$ such that total inputs and outputs stay feasible; i.e., we do not use more aggregated input, $x^{*1} + x^{*2} + x^{*3} \leq x^1 + x^2 + x^3$, and we produce at least the same aggregated output, $y^{*1} + y^{*2} + y^{*3} \geq y^1 + y^2 + y^3$. All of the new production plans must be feasible $(x^{*k}, y^{*k}) \in T$ for all $k = 1, 2, 3$. The largest proportional savings on original input usage that we can achieve via such reallocation can be calculated by solving the following program:

$$\begin{aligned} \min_{H,(x^{*j},y^{*j}),j=1,2,3} \quad & H \\ \text{s.t.} \quad & H(x^1 + x^2 + x^3) \geq (x^{*1} + x^{*2} + x^{*3}), \\ & (y^1 + y^2 + y^3) \leq (y^{*1} + y^{*2} + y^{*3}), \\ & (x^{*j}, y^{*j}) \in T, j = 1, 2, 3. \end{aligned}$$

Therefore, if $H = 0.9$, this means that we can save 10% of all resources used in the three firms by simply moving production around to take advantage of best practices, economics of scale, and economies of scope. So far, we have not made any assumptions about the underlying technology set, T, but if we assume that it is convex, we can actually show that the saving factor H shown above can also be calculated via as simple Farrell input efficiency evaluation of the average firm

$$H = \min_H \{ H \mid (H\frac{1}{3}(x^1 + x^2 + x^3), \frac{1}{3}(y^1 + y^2 + y^3)) \in T \} .$$

Thus, to calculate H, we can simply form the average firm, i.e. a hypothetical firm using the average of all input vectors to produce the average of all output vectors. The Farrell efficiency of this entity is a measure of what can be gained by everyones adjusting to best practices and by reallocating production between the 3 firms. We shall investigate such programs and some useful variations in Chap. 9.

Numerical example

As an example of the reallocation issue, consider a case in which 3 firms have produced 1 output using 1 inputs. The production frontier is given by

$$y = \sqrt{x - 5} \text{ for } x \leq 5.$$

The observed input–output combinations are

$$\begin{aligned} (10, \sqrt{10 - 5}) &= (10, 2.23), \\ (20, \sqrt{20 - 5}) &= (20, 3.87), \\ (30, \sqrt{30 - 5}) &= (30, 5). \end{aligned}$$

We see that they all operate on the efficient frontier, i.e., on an individual basis they cannot improve. However, if they collaborate and share resources and obligations, they may be able to conserve some of their aggregated input and still produce the same aggregated output. Specifically, following the guidelines above, we can measure the Farrell efficiency of the average firm. The average firm has used (10+20+30)/3= 20 input units to produce (2.23+3.87+5)/3=3.70 output units. The minimal input necessary to produce output of 3.70 is $3.70^2 + 5 = 18.71$. The minimal share of the average input that suffices to produce the average output is therefore

$$H = \frac{18.71}{20} = 0.94.$$

This result shows that via reallocation, this small industry could save 6 % of input. The reason is quite obvious in this simple single–input, single–output case because there are disadvantages to being small based on fixed costs and disadvantages of being large because of diminishing retursn to scale. Therefore, it is more advantageous to operate average-size firms.

2.9 Choice between efficiency measures

The question naturally arises as to which of the many possible efficiency measures to choose. There are several both applied and theoretical aspects of this.

One very important aspect is *controllability*. The inputs and outputs that can be controlled by the entities to be evaluated are important because it is generally not very informative or motivating to be judged on the basis of factors that you cannot control. Therefore, the choice between input- and output-based evaluations, between general evaluations or conditional evaluations where some factors are fixed, and between allocative and technical efficiency depends very much on controllability.

The time perspective is relevant because in the long run, more factors are usually variable. The level in a hierarchy that is evaluated is relevant. A divisional manager may, for example, be evaluated for technical efficiency, while an officer at the headquarters who is responsible for resource allocation may be more correctly evaluated based on allocative efficiency or, if prices are not available, using structural efficiency measures. A hospital may not have much control over demand, and as a result, input-based evaluations may be more relevant, while a farmer may have many fixed resources (land, etc.) and, therefore, should be evaluated more in terms of the output.

More generally, the *intended use* of the efficiency score is crucial. In a learning experience, the exact efficiency measurement is less important than the ability to find relevant peers taking into account the firms own preference, strategies, etc. The directional distance function approach may be particularly useful here due to its flexibility. In an allocation application, the distinction between fixed and variable inputs and outputs is often important, which might lead us to favor a Farrell approach, with some inputs and outputs that are non-discretionary (or even a directional distance function approach). In an incentive application, the task is to find an aggregation of performance that allows optimal contracting. We will see in later chapters that one can actually provide incentive rationales for radial measures like the Farrell approach.

On a very specific level, *ease of interpretation* is also important. One of the advantages of the Farrell measure in applications is that it is very easy to interpret. One can come up with many more or less ingenious ranking systems, and those that do not perform well may have very strong objections as to how the ranking was constructed and how the different performance dimensions were aggregated and weighted. One important element of the Farrell measure, however, is that it does not weigh the different dimensions. If a firm is not performing well according to this measure, it is very difficult for that firm to explain away the results because it is underperforming in all areas rather than just in one potentially overrated dimension. This is because the Farrell measure uses proportional changes. This argument can actually be given a game theoretical formalization, as we will show in Chap. 5.

As a last practical concern, let us mention *data availability* and *computations ease*. The more we know about values (prices, priorities), the more focused the evaluations can become. Prices for inputs, for example, enable us to conduct cost efficiency analyses that decompose efficiency into allocative and technical efficiency, which will provide us with more information han a pure technical efficiency analysis would. Likewise, using data from several years allows more robust evaluations and may possibly allow us to separately consider general productivity shifts and catch-up effects. Additionally, in more advanced applications involving, for example, complicated structural and network models, computational issues shall be considered. It is less interesting to dream up complicated calculations if they are very difficult to implement because the resulting programs become too non-linear, for example.

From a more theoretical perspective, we may compare the general properties of different measures using *axiomatic theory*. Some key results are given in Sect. 2.12. As emphasize there, the Farrell measure has several advantages but suffers from one problem: a lack of what is called indication. A firm may be efficient in the Farrell sense even if it is in fact not fully (Koopmans) efficient.

It is also important to keep the rational ideal model in mind when considering indices of technical efficiency. Ideally, efficiency should reflect utility effectiveness because efficiency is a sort of proxy for *utility effectiveness*. We know that dominance relationships are maintained under utility effectiveness in the sense that if one firm dominates another, then it is also more utility effective. We cannot, however, be sure that inefficient firms are less utility-effective then some efficient ones. Therefore, although efficiency provides a useful filter, efficiency is not a sufficient condition for firm effectiveness, and one should not be too fixated on the ability to make efficiency evaluations based on a minimum of assumptions. It is still important to think of ways to elicit preferences and make evaluations that more closely capture our preferences. After all, small improvements of the right type may be more valuable than large improvements to less important aspects.

2.10 Summary

In this chapter, we have taken a somewhat closer look at the general problem of evaluating and quantifying the performance of a firm by gauging it against a technology. We have defined efficiency as using the least resources to produce the most services, and we have looked at different ways to measure efficiency levels. We have covered the most widely used measure, the Farrell efficiency measure focusing on proportional improvements to inputs or outputs, and we have discussed alternative approaches like directional distance functions with excess, an additive measure of the number of times a given improvement bundle is feasible. We have also discussed how preference or price information allows more informative evaluations, including decompositions spotlighting allocative and technical efficiency factors. We have shown how one can distinguish between frontier shifts and catching up in a dynamic context and how structural efficiency can be evaluated by looking at networks of firms. Lastly, we have discussed some key concerns related to the choice between alternative measures. Some more advanced material, including the axiomatic characterization of some classical measures, is provided in Sect. 2.12 below.

2.11 Bibliographic notes

The notion of efficiency is used throughout economics and is perhaps most well-known in the context of the Pareto efficiency concept, wherein the outcomes for several individuals are compared using the efficiency criterion. A solution Pareto dominates another if, and only if, it makes someone better off without making anyone worse off. In multiple criteria decision-making, a main theme is how to find and choose among efficient alternatives, c.f. e.g., Bogetoft and Pruzan (1991). In a production economics context, the traditional reference is Koopmans (1951). The idea behind all related concepts is the same, however: we avoid weighing different persons, different criteria or different inputs and outputs together by using a more is better than less approach and looking for improvements that occur in some area without creating worse performance in others. In Bogetoft and Pruzan (1991), appendix 1, we formalize how efficiency is related to the rational ideal evaluations that economists seek to make.

The focus on proportional improvements was suggested by Debreu (1951) and Farrell (1957). The inverse of Farrell, the Shephard distance function, is due to Shephard (1953, 1970). The use of discretionary and non-discretionary dimensions is described in many textbooks: for example, Charnes et al (1995). However, this use dates back at least to Banker and Morey (1986).

The graph hyperbolic efficiency measure was suggested in Färe et al (1985), while basic work on the excess function was done by Luenberger (1992) and Chambers et al (1998). The idea of constructing interactive benchmarking systems was suggested in Bogetoft and Nielsen (2005) and Bogetoft et al (2006a) and commercialized in the Interactive Benchmarking IB^{TM} software from www.ibensoft.com used by Danish Waterworks.

The idea of allocative efficiency dates back to at least Debreu (1951) and Farrell (1957), while the Malmquist index dates back to Malmquist (1953) and was made popular by Caves et al (1982) and Färe et al (1994). There is a large body of literature on alternative modes of decomposition. Bogetoft et al (2006b) provides an alternative definition of allocative efficiency that allows us to calculate allocative efficiency without assuming that technical efficiency has first been eliminated.

The idea of structural efficiency dates back to at least Farrell (1957) on p.262. He defined structural efficiency as "the extent to which an industry keeps up with the performance of its own best firms" and suggested that it can be measured by comparing the horizontal aggregation of the industry's firms with the frontier constructed from its individual firms. A related approach is the average unit approach suggested by Försund and Hjalmarsson (1979). In a recent study, Andersen and Bogetoft (2007) developed a DEA-based reallocation model to study the potential gains from redistributing fishery quotas among Danish vessels. An interesting result was that the redistribution of production might be just as useful as the learning of best practices. This is relevant because it may be optimistic to suppose that all units can adopt best practices, at least in the short run, and reallocations off the frontier should therefore be considered, cf. also Bogetoft et al (2006b) The idea of interpreting this result as the possible effect of a reallocation program calculating H comes from Bogetoft and Wang (2005). The application for merger control is developed in Bogetoft and Katona (2008), while the application for the reallocation of agricultural production is described in Andersen and Bogetoft (2007) and Bogetoft et al (2009). We discuss structural efficiency and network models in more detail in Chap. 9, where we provide more references.

The link between efficiency and decision theory formalized in the appendix builds directly on Theorem 1 in Bogetoft and Pruzan (1991), where a proof is also provided.The axiomatic approach to efficiency evaluations was initiated by Färe and Lovell (1978). They worked with axioms 2, 3 and 4 below. This was followed by work by Russell (1985, 1987, 1990), Zieschang (1984), and others. Axiomatic characterizations of special directional distance measures and discussions of their relationship to bargaining theory are given in Bogetoft and Hougaard (1999).

2.12 Appendix: More advanced material on efficiency measures

As an appendix to this chapter on efficiency measures, we will now present some more technical material that can be skipped during a first reading.

2.12.1 The rationale of efficiency

It is of course possible to identify more precise and profound motivations for reliance on efficiency. To consider one such motivation, we will now look at efficiency in a decision theoretical context.

The basic economic model of (individual) choice is the *rational ideal model*. The rational ideal model depicts an economic entity (an individual or system) as seeking the best means to his desired ends; it is defined by the set of alternatives available and ones preferences regarding them.

Let us assume that a firm has transformed m inputs $x^* \in \mathbb{R}^m$into n outputs $y^* \in \mathbb{R}^n$. Additionally, let the objective or preference function be given by

$$U : \mathbb{R}^{m+n} \to \mathbb{R}$$

where $U(x, y)$ is the utility attached to a production plan (x, y) Also, let us assume that the set of feasible input-output vectors is

$$T \subseteq R^{m+n}.$$

In this set-up, we have that (x^*, y^*) is *optimal* if and only if it solves the *basic decision problem*

$$\begin{aligned} &\max\ U(x, y) \\ &s.t.\ \ (x, y) \in T \end{aligned} \tag{2.1}$$

i.e., if and only if the firm has made the best, most effective use of its potential.

In practice, this ideal evaluation can seldom be conducted. A common obstacle is that the feasible production plans T are not known. Another is that the firm or the evaluator may not have clear-cut expressions of the aggregate performance evaluation criterion $U(.)$. In public sector contexts, for example, where the agent could be a school, court or police station, it is often hard to imagine explicit delineations of the production options. Additionally, the multiple outputs produced will be difficult to aggregate, such that an explicit preference structure is usually not available.

One perspective on the modern theory of productivity analysis is that it allows us to make evaluations in contexts with incomplete information about options T and preferences $U(.)$. This is done by focusing on efficiency instead of effectiveness and by focusing on efficiency relative to a constructed technology rather than in relation to the underlying true but unknown technology.

Let us now focus on the first problem, the lack of information about $U(x, y)$ and the resulting need to shift our attention from effectiveness to efficiency.

As a matter of notation, recall that an input–output combination $(x', y') \in T \subseteq R^{m+n}$ is efficient relative to the technology T if and only if

$$\forall (x, y) \in T : x \leq x', y \geq y' \Rightarrow x = x', y = y'.$$

The set of *efficient* plans is denoted T^E.

The focus on efficiency is natural. On the one hand, efficiency is not too strong a requirement because one can always find an optimal production plan among the efficient ones, and on the other hand, we cannot strengthen the efficiency requirement because any efficient plan may be the uniquely optimal plan based on one of the underlying but unknown preference functions.

We formalize these reflections in the following proposition.

Proposition of Rational Efficiency. *For a basic decision problem (2.1) where U is weakly increasing in y and weakly decreasing in x, i.e., $x \leq x', y \geq y' \Rightarrow U(x, y) \geq U(x', y')$, we have that*

1. *for any (x^*, y^*) optimal in (2.1), there exists a $(x, y) \in T^E$ such that $U(x, y) = U(x^*, y^*)$, and*
2. *for any $(x^*, y^*) \in T^E$, there exists a U such that (x^*, y^*) is a unique solution to (2.1).*

This is a straightforward modification of a well-known result in decision theory.

According to the proposition, we do not lose anything by focusing on efficient production plans. By the first bullet, an optimal alternative can always be found among the set of efficient alternatives. However, the set that we consider to find the optimal alternative cannot a priori be a smaller set than T^E if all we know about the preference function or the overall evaluation criteria U is that they are weakly increasing. By the second bullet, any efficient alternative may turn out to be the only optimal plan for a weakly increasing U. Thus, *the efficient set T^E is the smallest sufficient set of alternatives to consider.*

As noted in Sect. 2.4 and below in the axiomatic characterization, Farrell efficiency does not guarantee efficiency because there may be slack left when we project a point onto the frontier of the technology. This should not, however, disturb us too much. After all, when we use the radial measures, we simply find more Farrell-efficient points than truly efficient points, i.e. we do not exclude any interesting points a priori, but we may leave uninteresting points in the Farrell efficient set.

2.12.2 Axiomatic characterization of efficiency measures

To understand the pros and cons of different benchmarking approaches, it is useful to develop a basic understanding of the properties of the efficiency measures that we use. Here we introduce some desirable properties of efficiency measures and then record which of them the Farrell measures (and a few other measures) have.

To simplify the exposition, we focus on the input space. The technology can therefore be defined as the input set L, i.e. the set of input combinations $x \in \mathbb{R}^m_+$ that can produce a fixed amount of output $y \in \mathbb{R}^n_+$. Formally, $L(y) = \{ y \in \mathbb{R}^m_+ \mid (x, y) \in T \}$ and to simplify we just write L as y is fixed in what follows. With standard regularity assumptions on T if follows that $L \subset \mathbb{R}^m_+$ has the properties of being non-empty, closed and free disposable, and every $x \in L$ can produce y. Define the weakly efficient (Farrell efficient) subset of L (i.e., the isoquant of L,) as $I = \{x \in L \mid \lambda \in [0,1[\Rightarrow \lambda x \notin L\}$, and the efficient (Koopmans efficient) subset of L as $L^E = \left\{x \in L \middle| \forall x' \in \mathbb{R}_+^m : x' \leq x \wedge x' \neq x \Rightarrow x' \notin L\right\}$, in two dimensions the part of the isoquant that does not contain vertical or horizontal parts.

To be of any general interest, an efficiency concept must be applicable to a reasonably large class of technologies: for example, any technology in a set $\mathcal{L}$ with the properties listed for L above. Note that an efficiency measure basically maps a production plan and a technology into the real numbers. We can formally define it in the following way. *An efficiency measure or index* is a function

$$\epsilon : \mathbb{R}_+^m \times \mathcal{L} \to \mathbb{R}$$

such that $\epsilon(x, L) \in [0, 1]$ for $x \in L$.

We see that the Farrell efficiency measure satisfies these conditions.

Another measure that has been around for several years is the *Färe-Lovell efficiency* index.

$$E_{FL}(x, L) =$$

$$\min\Big\{ \frac{1}{\#\{i \,|\, x_i > 0\}} \sum_{i=1}^{m} \lambda_i \;\Big|\; (\lambda_1 x_1, \ldots, \lambda_m x_m) \in L,\ \lambda_i \in [0,1],\ x \in L \Big\}.$$

The idea of this measure is that we try to minimize the average of the input-specific contraction factors; i.e. we conduct individual contractions of the different inputs. Hence, this process does not necessarily lead to proportional reductions as in the Farrell case. Graphically, and presuming that all $x_i > 0$, $i = 1, \ldots, m$, the measure corresponds to comparing x to the point on L that minimizes a cost function with prices $(x_1^{-1}, x_2^{-1}, \ldots, x_m^{-1})$. The reason is that minimizing $\sum_{i=1}^m \lambda_i$ under the restriction $(\lambda_1 x_1, \ldots, \lambda_m x_m) \in L$ is equivalent to minimizing $\sum_{i=1}^m x_i^{-1} \tilde{x}_i$ over $\tilde{x} \leq x, \tilde{x} \in L$. Simply substitute using $\lambda_i x_i = \tilde{x}_i$.

A third measure combining the two is the *Zieschang index* defined as

$$E_Z(x, L) = E(x, L) E_{FL}(Ex, L)$$

which corresponds to Farrell efficiency multiplied by the Färe-Lovell efficiency of the Farrell projected input combination.

It is easy to see that

$$E(x, L) \geq E_Z(x, L) \geq E_{FL}(x, L).$$

Now let us consider some general and desirable properties:

Commensurability / Invariance to permutations and rescaling (A1) For all $m \times m$ matrices with exactly one non-zero and positive element in each row and column, we have that $\epsilon(x, L) = \epsilon(Ax, AL)$.

Indication (A2) $\epsilon(x, L) = 1$ if and only if $x \in L^E$.

Homogeneity of degree -1 *(A3)* $\epsilon(\lambda x, L) = \lambda^-1\epsilon(x, L)$ for all λ where $\lambda x \in L$.
Monotonicity in inputs (A4) $x' \geq x, x' \neq x$ implies $\epsilon(x, L) > \epsilon(x', L)$.
Continuity (A5) $\epsilon(\lambda x, L_e)$ is a continuous function of $\lambda \geq 1$ when $L_e = \{x \in \mathbb{R}^m_+ \mid x \geq e\}$ and $x \in L_e$.

We interpret these properties as follows:

Commensurability (A1) means that efficiency is not affected by different permutations of inputs, i.e. it does not matter in what order we list the inputs. Moreover, efficiency is independent of linear re-scalings of the different inputs. Thus, for example, it does not matter if we measure in kg or tons. Both E, E_{FL} and E_Z clearly have this property.

Indication (A2) means that we only assign the value 1 to points that are efficient (in the Koopmanns sense). The Farrell measure does not have this property because the radial project may end at a vertical or horizontal part of the isoquant. This is the major drawback of the Farrell measure and one of the motivations for the Färe-Lovell and Zieschang indices.

Homogeneity of degree -1 (A3) means that if we double the inputs, we halve the efficiency. Farrell efficiency and Zieschang efficiency satisfy this, but the Färe-Lovell's index does not.

Monotonicity of inputs (A4) requires that if we increase the usage of at least one input, we lower the efficiency score.

Continuity (A5) requires that if we have a Leontief technology and vary the input consumption proportionally for the technology, then the efficiency score will vary continuously. This is a desirable property in practice because we do not want small data errors to have dramatic impact on the efficiency score. Unfortunately, this requirement is not easy to fulfill.

Although these properties seem reasonable, they are not easy to fulfill. As a general *non-existence theorem*, we note that one cannot construct a measure that satisfies A2, A3 and A4 simultaneously for the ample class of technologies, $\mathcal{L}$, that we have considered here. In this sense there is no best efficiency measure to always be used in efficiency and benchmark analysis.

Table 2.4 summarizes the properties of the Farrell, Färe-Lovell and Zieschang measures. In the table, (Yes) means Yes as long as we only consider strictly positive input vectors.

Table 2.4 Properties of efficiency measures

Property	E	E_{FL}	E_Z
Commensurability (A1)	Yes	Yes	Yes
Indication A2	No	Yes	Yes
Homogeneity A3	Yes	No	Yes
Monotonicity A4	No	(Yes)	No
Continuity A5	(Yes)	No	No