

Chapter 2

THE MEASUREMENT OF EFFICIENCY

This chapter is about the measurement of efficiency of production units. It opens with a section concerning basic definitions of productivity and efficiency. After that, an historical and background section follows, reporting some of the most important contributions until around '90s. Then, the axiomatic underpinning of the Activity Analysis framework used to represent the production process is described in the economic model section. Afterwards, efficient frontier models are classified according to three main *criteria*: specification (or not) of the form of the frontier; presence of noise in the estimation procedure; type of data analyzed (cross-section or panel data). The presentation of the most known nonparametric estimators of frontiers (*i.e.*, Data Envelopment Analysis (DEA) and Free Disposal Hull (FDH)) is subsequently. Finally, a section summarizing recent developments in nonparametric efficiency analysis concludes the chapter.

2.1 Productivity and Efficiency

According to a classic definition (see *e.g.* Vincent 1968) *productivity* is the *ratio* between an output and the factors that made it possible. In the same way, Lovell (1993) defines the *productivity* of a production unit as the ratio of its output to its input.

This ratio is easy to compute if the unit uses a single input to produce a single output. On the contrary, if the production unit uses several inputs to produce several outputs, then the inputs and outputs have to be aggregated so that productivity remains the ratio of two scalars.

We can distinguish between a *partial* productivity, when it concerns a sole production factor, and a *total factor* (or global) productivity, when referred to all (every) factors.

Similar, but not equal, is the concept of efficiency. Even though, in the efficiency literature many authors do not make any difference between productivity and efficiency. For instance, Sengupta (1995) and Cooper, Seiford and Tone (2000) define both productivity and efficiency as the ratio between output and input.

Instead of defining the efficiency as the ratio between outputs and inputs, we can describe it as a distance between the quantity of input and output, and the quantity of input and output that defines a frontier, the best possible frontier for a firm in its cluster (industry).

Efficiency and productivity, anyway, are two cooperating concepts. The measures of efficiency are more accurate than those of productivity in the sense that they involve a comparison with the most efficient frontier, and for that they can complete those of productivity, based on the ratio of outputs on inputs.

Lovell (1993) defines the efficiency of a production unit in terms of a comparison between observed and optimal values of its output and input. The comparison can take the form of the ratio of observed to maximum potential output obtainable from the given input, or the ratio of minimum potential to observed input required to produce the given output. In these two comparisons the optimum is defined in terms of production possibilities, and efficiency is technical.

Koopmans (1951; p. 60) provided a definition of what we refer to as *technical efficiency*: an input-output vector is technically efficient if, and only if, increasing any output or decreasing any input is possible only by decreasing some other output or increasing some other input.

Farrell (1957; p. 255) and much later Charnes and Cooper (1985; p. 72) go back over the empirical necessity of treating Koopmans' definition of technical efficiency as a relative notion, a notion that is relative to best observed practice in the reference set or comparison group. This provides a way of differentiating efficient from inefficient production units, but it offers no guidance concerning either the degree of inefficiency of an inefficient vector or the identification of an efficient vector or combination of efficient vectors against which comparing an inefficient vector.

Debreu (1951) offered the first measure of *productive efficiency* with his *coefficient of resource utilization*. Debreu's measure is a radial measure of technical efficiency. Radial measures focus on the maximum feasible *equiproportionate* reduction in all variable inputs, or the maximum feasible *equiproportionate* expansion of all outputs. They are independent of unit of measurement.

Applying radial measures the achievement of the maximum feasible input contraction or output expansion suggests technical efficiency, even though there may remain *slacks* in inputs or *surpluses* in output. In economics the notion of efficiency is related to the concept of Pareto optimality. An input-output bundle is not Pareto optimal if there remains the opportunity of any net increase in

outputs or decrease in inputs. Pareto-Koopmans measures of efficiency (*i.e.*, measures which call a vector efficient if and only if it satisfies the Koopmans definition reported above, coherent with the Pareto optimality concept) have been analysed in literature. See *e.g.*, Färe (1975), Färe and Lovell (1978) and Russell (1985, 1988, 1990) among others.

Farrell (1957) extended the work initiated by Koopmans and Debreu by noting that production efficiency has a second component reflecting the ability of producers to select the “right” technically efficient input-output vector in light of prevailing input and output prices. This led Farrell to define overall productive efficiency as the product of *technical* and *allocative* efficiency. Implicit in the notion of allocative efficiency is a specific behavioral assumption about the goal of the producer; Farrell considered cost-minimization in competitive inputs markets, although all the behavioral assumptions can be considered. Although the natural focus of most economists is on markets and their prices and thus on allocative rather than technical efficiency and its measurement, he expressed a concern about human ability to measure prices accurately enough to make good use of allocative efficiency measurement, and hence of overall economic efficiency measurement. This worry expressed by Farrell (1957; p. 261) has greatly influenced the OR/MS work on efficiency measurement. Charnes and Cooper (1985; p. 94) cite Farrell concern as one of several motivations for the typical OR/MS emphasis on the measurement of technical efficiency.

It is possible to distinguish different kind of efficiency, such as scale, allocative and structural efficiency.

The *scale efficiency* has been developed in three different ways. Farrell (1957) used the most restrictive technology having constant returns to scale (CRS) and exhibiting strong disposability of inputs. This model has been developed in a linear programming framework by Charnes, Cooper and Rhodes (1978). Banker, Charnes and Cooper (1984) have shown that the CRS measure of efficiency can be expressed as the product of a technical efficiency measure and a scale efficiency measure. A third method of scale uses nonlinear specification of the production function such as Cobb-Douglas or a translog function, from which the scale measure can be directly computed (see Sengupta, 1994 for more details).

The *allocative efficiency* in economic theory measures a firm’s success in choosing an optimal set of inputs with a given set of input prices; this is distinguished from the technical efficiency concept associated with the production frontier, which measures the firm’s success in producing maximum output from a given set of inputs.

The concept of *structural efficiency* is an industry level concept due to Farrell (1957), which broadly measures in what extent an industry keeps up with the performance of its own best practice firms; thus it is a measure at the industry level of the extent to which its firms are of optimum size *i.e.* the extent to which

the industry production level is optimally allocated between the firms in the short run. A broad interpretation of Farrell's notion of structural efficiency can be stated as follows: industry or cluster A is more efficient structurally than industry B, if the distribution of its best firms is more concentrated near its efficient frontier for industry A than for B. In their empirical study, Bjurek, Hjalmarsson and Forsund (1990) compute structural efficiency by simply constructing an average unit for the whole cluster and then estimating the individual measure of technical efficiency for this average unit. On more general aggregation issues, see Färe and Zelenyuk (2003) and Färe and Grosskopf (2004, p. 94 ff).

2.2 A short history of thought

The theme of productive efficiency has been analysed since Adam Smith's pin factory and before¹. However, as we have seen in the previous section, a rigorous analytical approach to the measurement of efficiency in production originated only with the work of Koopmans (1951) and Debreu (1951), empirically applied by Farrell (1957).

An important contribution to the development of efficiency and productivity analysis has been done by Shephard's models of technology and his distance functions (Shephard 1953, 1970, 1974). In contrast to the traditional production function, direct input and output correspondences admit multiple outputs and multiple inputs. They are thus able to characterize all kinds of technologies without unwarranted output aggregation prior to analysis. The Shephard direct input distance function treats multiple outputs as given and contracts inputs vectors as much as possible consistent with technological feasibility of contracted input vector. Among its several useful properties, one of the most important is the fact that the *reciprocal* of the direct input distance function has been proposed by Debreu (1951) as a coefficient of resource utilization, and by Farrell (1957) as a measure of technical efficiency. This property has both a theoretical and a practical significance. It allows the direct input distance function to serve two important roles, simultaneously. It provides a complete characterization of the structure of multi-input, multi-output efficient production technology, and a reciprocal measure of the distance from each producer to that efficient technology.

The main role played by the direct input distance function is to gauge technical efficiency. Nevertheless, it can also be used to construct input quantity indexes (Tornqvist, 1936; Malmquist, 1953) and productivity indexes (Caves, Christensen, and Diewert, 1982). Similarly, the direct output distance function introduced by Shephard (1970) and the two indirect distance functions of Shephard (1974) can be used to characterize the structure of efficient production

¹This section is based on Färe, Grosskopf and Lovell (1994), pp. 1-23; and Kumbhakar and Lovell (2000), pp. 5-7.

technology in the multi-product case, to measure efficiency to that technology, and to construct output quantity indexes (Bergson, 1961; Moorseen, 1961) and productivity indexes (Färe, Grosskopf, and Lovell, 1992).

Linear programming theory is a milestone of efficiency analysis. The work of Dantzig (1963) is closely associated with linear programming since he contributed to the basic computational algorithm (the simplex method) used to solve this problem. Charnes and Cooper (1961) made considerable contributions to both theory and application in the development of linear programming, and popularize its application in DEA in the late 70s (see Charnes, Cooper and Rhodes, 1978). Forsund and Sarafoglou (2002) offer an interesting historical reconstruction of the literature developments subsequent to Farrell's seminal paper that lead to the introduction of the DEA methodology.

The use of linear programming and activity analysis can be found in the work of Leontief (1941, 1953) who developed a special case of activity analysis which has come to be known as input-output analysis. Whereas Leontief's work was directed toward constructing a workable model of general equilibrium, efficiency and productivity analysis is more closely related to the microeconomic production programming models developed by Shephard (1953, 1970, 1974), Koopmans (1951, 1957) and Afriat (1972). In these models observed activities, such as the inputs and outputs of some production units, serve as coefficients of activity or intensity variables forming a series of linear inequalities, yielding a piecewise linear frontier technology.

The work of Koopmans and Shephard imposes convexity on the reference technology, therefore, the DEA estimator relies on the convexity assumption. The Free Disposal Hull (FDH) estimator, that maintains free disposability while relaxes convexity, was introduced by Deprins, Simar and Tulkens (1984).

By enveloping data points with linear segments, the programming approach reveals the structure of frontier technology without imposing a specific functional form on either technology or deviations from it.

Frontier technology provides a simple means of computing the distance to the frontier - as a maximum feasible radial contraction or expansion of an observed activity. This means of measuring the distance to the frontier yields an interpretation of performance or efficiency as maximal-minimal proportionate feasible changes in an activity given technology. This explanation is consistent with Debreu's (1951) coefficient of resource utilization and with Farrell's (1957) efficiency measures. However, neither Debreu nor Farrell formulated the efficiency measurement problem as a linear programming problem, even though Farrell and Fieldhouse (1962) envisaged the role of linear programming. The full development of linear programming techniques took place later. Boles (1966), Bressler (1966), Seitz (1966) and Sitorius (1966) developed the piecewise linear case, and Timmer (1971) extended the piecewise log-linear case.

Linear programming techniques are also used in production analysis for non-parametric ‘tests’² on regularity conditions and behavioral objectives. Afriat (1972) developed a series of consistency ‘tests’ on production data by assuming an increasing number of more restrictive regularity hypotheses on production technology. In so doing he expanded his previous work on utility functions (Afriat 1967) based on the revealed preference analysis (Samuelson, 1948).

These ‘tests’ of consistency, as well as similar ‘tests’ of hypotheses proposed by Hanoch and Rothschild (1972), are all based on linear programming formulations. Diewert and Parkan (1983) suggested that this battery of tools could be used as a screening device to construct frontiers and measure efficiency of data relative to the constructed frontiers. Varian (1984, 1985, 1990) and Banker and Maindiratta (1988) extended the Diewert and Parkan approach. In particular, Varian seeks to reduce the “all-or-nothing” nature of the tests - either data pass a test or they do not - by developing a framework for allowing small failures to be attributed to measurement in the data rather than to failure of the hypothesis under investigation.

All these studies use nonparametric linear programming models to explore the consistency of a dataset, or a subset of a dataset, with a structural (*e.g.* constant return to scale) or parametric (*e.g.* Cobb-Douglas) or behavioral (*e.g.* cost minimization) hypothesis. These tools, originally proposed as screening devices to check for data accuracy, provide also guidance in the selection of parametric functional forms as well as procedures useful to construct frontiers and measure efficiency. The problem of nonparametric exploration of regularity conditions and behavioral objectives has been treated also by Chavas and Cox (1988, 1990), Ray (1991), and Ray and Bhadra (1993).

Some works have *indirectly* influenced the development of the efficiency and productivity analysis. Hicks (1935, p.8) states his “easy life” hypothesis as follows: “people in monopolistic positions [...] are likely to exploit their advantage much more by not bothering to get very near the position of maximum profit, than by straining themselves to get very close to it. The best of all monopoly profits is a quiet life”. The suggestion of Hicks, *i.e.* the fact that the absence of competitive pressure might allow producers the freedom to not fully optimize conventional objectives, and, by implication, that the presence of competitive pressure might force producers to do so, has been adopted by many authors (see *e.g.* Alchian and Kessel, 1962, and Williamson, 1964).

Another field of work, related to efficiency literature, is the property rights field of research, which asserts that public production is inherently less efficient than private production. This argument, due originally to Alchian (1965), states that concentration and transferability of private ownership shares create

²Here and below when we use the word test between quotation mark we mean qualitative indicators that are not real statistical test procedures.

an incentive for private owners to monitor managerial performance, and that this incentive is diminished for public owners, who are dispersed and whose ownership is not transferable. Consequently, public managers have wider freedom to pursue their own at the expense of conventional goals. Thus Niskanen (1971) argued that public managers are budget maximizers, de Alessi (1974) argued that public managers exhibit a bias toward capital-intensive budgets, and Lindsay (1976) argued that public managers exhibit a bias toward “visible” inputs. However, ownership forms are more varied than just private or public. Hansmann (1988), in fact, identifies investor-owned firms, customer-owned firms, worker-owned firms, as well as firms without owners (nonprofit enterprisers). Each of them deals in a different way with problems associated with hierarchy, coordination, incomplete contracts and monitoring and agency costs. This leads to the expectation that different ownership forms will generate differences in performance.³

As a more micro level is concerned, Simon (1955, 1957) analyzed the performance of producers in the presence of bounded rationality and satisfying behavior. Later Leibenstein (1966, 1975, 1976, 1978, 1987) argued that production is bound to be inefficient as a result of motivation, information, monitoring, and agency problems within the firm. This type of inefficiency, the so called “X-inefficiency” has been criticized by Stigler (1976) and de Alessi (1983) among others since it reflects an incompletely specified model rather than a failure to optimize.

The problem of model specification - including a complete list of inputs and outputs, and perhaps conditioning variables as well, a list of constraints, technological, and other (*e.g.* regulatory) is a difficult issue to face. Among others, Banker, Chang and Cooper (1996) analyse the effects of misspecified variables in DEA. Simar and Wilson (2001) propose a statistical procedure to test for the relevance of inputs/outputs in DEA models.

This literature suggests that the development of efficiency analysis is particularly useful if and when it could be used to shed empirical light on the theoretical issues outlined above.

2.3 The economic model

In this paragraph we describe the main axioms on which the economic model underlined the measurement of efficiency is based on.⁴

Much empirical evidence suggests that although producers may indeed attempt to optimize, they do not always succeed. Not all producers are always so successful in solving their optimization problems. Not all producers succeed

³This expectation is based on a rich theoretical literature. See *e.g.* the “classical” survey by Holmstrom and Tirole (1989).

⁴See also Färe and Grosskopf (2004), pp.151-161.

in utilizing the minimum inputs required to produce the outputs they choose to produce, given the technology at their disposal. In light of the evident failure of at least some producers to optimize, it is desirable to recast the analysis of production away from the traditional production function approach toward a frontier based approach. Hence we are concerned with the estimation of frontiers, which envelop data, rather than with functions, which intersect data.

In this setting, the main purpose of productivity analysis studies is to evaluate numerically the performance of a certain number of firms (or business units or Decision Making Units, DMU) from the point of view of *technical efficiency*, *i.e.* their ability to operate close to, or on the boundary of their production set. The problem to be analyzed is thus set in terms of physical input and output quantities.

We assume to have data in cross-sectional form, and for each firm we have the value of its inputs and outputs used in the production process. Measuring efficiency for any data set of this kind requires first to determine what the boundary of the production set can be; and then to measure the distance between any observed point and the boundary of the production set.

Given a list of p inputs and q outputs, in economic analysis the operations of any productive organization can be defined by means of a set of points, Ψ , the *production set*, defined as follows in the Euclidean space \mathcal{R}_+^{p+q} :

$$\Psi = \{(x, y) \mid x \in \mathcal{R}_+^p, y \in \mathcal{R}_+^q, (x, y) \text{ is feasible}\}, \quad (2.1)$$

where x is the input vector, y is the output vector and “feasibility” of the vector (x, y) means that, within the organization under consideration, it is physically possible to obtain the output quantities y_1, \dots, y_q when the input quantities x_1, \dots, x_p are being used (all quantities being measured per unit of time). It is useful to define the set Ψ in terms of its *sections*, defined as the images of a relation between the input and the output vectors that are the elements of Ψ . We can define then the *input requirement set* (for all $y \in \Psi$) as:

$$C(y) = \{x \in \mathcal{R}_+^p \mid (x, y) \in \Psi\}. \quad (2.2)$$

An input requirement set $C(y)$ consists of all input vectors that can produce the output vector $y \in \mathcal{R}_+^q$.

The *output correspondence set* (for all $x \in \Psi$) can be defined as:

$$P(x) = \{y \in \mathcal{R}_+^q \mid (x, y) \in \Psi\}. \quad (2.3)$$

$P(x)$ consists of all output vectors that can be produced by a given input vector $x \in \mathcal{R}_+^p$.

The production set Ψ can also be retrieved from the inputs sets, specifically:

$$\Psi = \{(x, y) \mid x \in C(y), y \in \mathcal{R}_+^q\}. \quad (2.4)$$

Furthermore, it holds that:

$$(x, y) \in \Psi \Leftrightarrow x \in C(y), y \in P(x), \quad (2.5)$$

which tells us that the output and input sets are equivalent representations of the technology, as is Ψ .

The isoquants or efficient boundaries of the sections of Ψ can be defined in radial terms (Farrell, 1957) as follows. In the input space:

$$\partial C(y) = \{x | x \in C(y), \theta x \notin C(y), \forall \theta, 0 < \theta < 1\} \quad (2.6)$$

and in the output space:

$$\partial P(x) = \{y | y \in P(x), \lambda y \notin P(x), \forall \lambda > 1\}. \quad (2.7)$$

The axiomatic approach to production theory (Activity Analysis framework) assumes that the technology (production model) satisfies certain properties or axioms. These properties can be equivalently stated on Ψ , $P(x)$, $x \in \mathcal{R}_+^p$, $C(y)$, $y \in \mathcal{R}_+^q$.

Some economic axioms (EA) are usually done in this framework (on these concepts see also Shephard, 1970).

EA1: NO FREE LUNCH. $(x, y) \notin \Psi$ if $x = 0, y \geq 0, y \neq 0$.⁵

This axiom states that inactivity is always possible, *i.e.*, zero output can be produced by any input vector $x \in \mathcal{R}_+^p$, but it is impossible to produce output without any inputs.

EA2: FREE DISPOSABILITY. Let $\tilde{x} \in \mathcal{R}_+^p$ and $\tilde{y} \in \mathcal{R}_+^q$, with $\tilde{x} \geq x$ and $\tilde{y} \leq y$, if $(x, y) \in \Psi$ then $(\tilde{x}, y) \in \Psi$ and $(x, \tilde{y}) \in \Psi$.

This is the *free disposability* assumption, named also the ‘possibility of destroying goods without costs’, on the production set Ψ .

The *free disposability* (also called *strong disposability*) of outputs can be stated as follows: $y_1 \in P(x), y_2 \leq y_1$ then $y_2 \in P(x)$ or equivalently $y_1 \leq y_2$ then $C(y_2) \subseteq C(y_1)$. The *free disposability* of inputs can be defined as below: $x_1 \in C(y), x_2 \geq x_1$ then $x_2 \in C(y)$ or equivalently $x_1 \leq x_2$ then $P(x_1) \subseteq P(x_2)$.

The free disposability of both inputs and outputs is as follows:

$$\forall (x, y) \in \Psi, \text{ if } x' \geq x \text{ and } y' \leq y \text{ then } (x', y') \in \Psi.$$

We have also a *weak* disposability of inputs and outputs:

⁵Here and throughout inequalities involving vectors are defined componentwise, *i.e.* on an element-by-element basis.

- Weak disposability of inputs:

$$x \in C(y) \Rightarrow \forall \alpha \geq 1, \alpha x \in C(y) \text{ or } P(x) \subseteq P(\alpha x);$$

- Weak disposability of outputs:

$$y \in P(x) \Rightarrow \forall \alpha \in [0, 1], \alpha y \in P(x) \text{ or } C(\alpha y) \subseteq C(y).$$

The weak disposability property allows us to model congestion and overutilization of inputs/outputs.

EA3: BOUNDED. $P(x)$ is bounded $\forall x \in \mathcal{R}_+^p$.

EA4: CLOSENESS. Ψ is closed, $P(x)$ is closed, $\forall x \in \mathcal{R}_+^p$, $C(y)$ is closed, $\forall y \in \mathcal{R}_+^q$.

EA5: CONVEXITY. Ψ is convex. The convexity of Ψ can be stated as follows:

If $(x_1, y_1), (x_2, y_2) \in \Psi$, then $\forall \alpha \in [0, 1]$ we have :

$$(x, y) = \alpha(x_1, y_1) + (1 - \alpha)(x_2, y_2) \in \Psi.$$

EA6: CONVEXITY OF THE REQUIREMENT SETS. $P(x)$ is convex $\forall x \in \mathcal{R}_+^p$ and $C(y)$ is convex $\forall y \in \mathcal{R}_+^q$.

If Ψ is convex, then the inputs and outputs sets are also convex, i.e. EA5 implies EA6.

A further characterization of the shape of the frontier relates to returns to scale (RTS). According to a standard definition in economics, RTS express the relation between a proportional change in inputs to a productive process and the resulting proportional change in output. If an n per cent rise in all inputs produces an n per cent increase in output, there are constant returns to scale (CRS). If output rises by a larger percentage than inputs, there are increasing returns to scale (IRS). If output rises by a smaller percentage than inputs, there are decreasing returns to scale (DRS). Returns to scale can be described as properties of the correspondence sets $C(y)$ and/or $P(x)$. We follow here the presentation of Simar and Wilson (2002, 2006b). The frontier exhibits *constant returns to scale* (CRS) everywhere if and only if:

$$\forall (x, y) \text{ s.t. } x \in \partial C(y) \text{ then } \alpha x \in \partial C(\alpha y), \forall \alpha > 0$$

or equivalently⁶,

$$\forall \alpha > 0, C(\alpha y) = \alpha C(y).$$

⁶Analogous expressions hold in terms of $P(x)$: $\forall \alpha > 0, P(\alpha x) = \alpha P(x)$.

Constant Returns to Scale in the *neighborhood* of a point (x, y) s.t. $x \in \partial C(y)$ are characterized by $C(\alpha y) = \alpha C(y)$ for some $\alpha > 0$.

Increasing Returns to Scale in the *neighborhood* of a point (x, y) s.t. $x \in \partial C(y)$ implies that $(\alpha x, \alpha y) \notin \Psi$ for $\alpha < 1$.

Decreasing Returns to Scale in the *neighborhood* of a point (x, y) s.t. $x \in \partial C(y)$ implies that $(\alpha x, \alpha y) \notin \Psi$ for $\alpha > 1$.

A frontier that exhibits increasing, constant and decreasing returns to scale in different regions is a *Variable Returns to Scale* (VRS) frontier.

The assumptions we have introduced here are intended to provide enough structure to create meaningful and useful technologies. Generally speaking, we will not impose all of these axioms on a particular technology, rather we will select subsets of these assumptions that are suitable for the particular problem under study.

Turning back to the production set itself, the above definitions allow us to characterize any point (x, y) in Ψ as:

input efficient if $x \in \partial C(y)$

input inefficient if $x \notin \partial C(y)$

output efficient if $y \in \partial P(x)$

output inefficient if $y \notin \partial P(x)$.

From what stated above, DMUs are efficient, *e.g.* in an input-oriented framework, if they are on the boundary of the input requirement set (or, for the output oriented case, on the boundary of the output correspondence set). In some cases, however, these efficient firms may not be using the fewest possible inputs to produce their outputs. This is the case where we have slacks. This is due to the fact that the Pareto-Koopmans efficient subsets of the boundaries of $C(y)$ and $P(x)$, *i.e.* $\text{eff } C(y)$ and $\text{eff } P(x)$, may not coincide with the Farrell-Debreu boundaries $\partial C(y)$ and $\partial P(x)$, *i.e.*⁷:

$$\text{eff } C(y) = \left\{ x \mid x \in C(y), x' \notin C(y) \forall x' \leq x, x' \neq x \right\} \subseteq \partial C(y), \quad (2.8)$$

$$\text{eff } P(x) = \left\{ y \mid y \in P(x), y' \notin P(x) \forall y' \geq y, y' \neq y \right\} \subseteq \partial P(x). \quad (2.9)$$

⁷We give an illustration in Section 2.5 in Figure 2.2 where we describe DEA estimators of efficient frontier.

Once the efficient subsets of Ψ have been defined, we may define the efficiency measure of a firm operating at the level (x_0, y_0) by considering the distance from this point to the frontier. There are several ways to achieve this but a simple way suggested by Farrell (1957), in the lines of Debreu (1951), is to use a *radial* distance from the point to its corresponding frontier. In the following we will concentrate our attention on radial measures of efficiency. Of course, we may look at the efficient frontier in two directions: either in the input direction (where the efficient subset is characterized by $\partial C(y)$) or in the output direction (where the efficient subset is characterized by $\partial P(x)$).

The Farrell input measure of efficiency for a firm operating at level (x_0, y_0) is defined as:

$$\theta(x_0, y_0) = \inf\{\theta | \theta x_0 \in C(y_0)\} = \inf\{\theta | (\theta x_0, y_0) \in \Psi\}, \quad (2.10)$$

and its Farrell output measure of efficiency is defined as:

$$\lambda(x_0, y_0) = \sup\{\lambda | \lambda y_0 \in P(x_0)\} = \sup\{\lambda | (x_0, \lambda y_0) \in \Psi\}. \quad (2.11)$$

So, $\theta(x_0, y_0) \leq 1$ is the radial contraction of inputs the firm should achieve to be considered as being input-efficient in the sense that $(\theta(x_0, y_0)x_0, y_0)$ is a frontier point. In the same way $\lambda(x_0, y_0) \geq 1$ is the proportionate increase of output the firm should achieve to be considered as being output efficient in the sense that $(x_0, \lambda(x_0, y_0)y_0)$ is on the frontier.

It is interesting to note that the efficient frontier of Ψ , in the radial sense, can be characterized as the units (x, y) such that $\theta(x, y) = 1$, in the input direction (belonging to $\partial C(y)$) and by the (x, y) such that $\lambda(x, y) = 1$, in the output direction (belonging to $\partial P(x)$). If the frontier is continuous, frontier points are such that $\theta(x, y) = \lambda(x, y) = 1$. The efficient frontier is unique but we have two ways to characterize it.

It is sometimes easier to measure these radial distances by their inverse, known as *Shephard distance functions* (Shephard, 1970). The Shephard input distance function provides a normalized measure of Euclidean distance from a point $(x, y) \in \mathcal{R}_+^{p+q}$ to the boundary of Ψ in a radial direction orthogonal to y and is defined as:

$$\delta^{in}(x, y) = \sup\{\theta > 0 | (\theta^{-1}x, y) \in \Psi\} \equiv (\theta(x, y))^{-1}, \quad (2.12)$$

with $\delta^{in}(x, y) \geq 1, \forall (x, y) \in \Psi$. Similarly, the Shephard output distance function provides a normalized measure of Euclidean distance from a point $(x, y) \in \mathcal{R}_+^{p+q}$ to the boundary of Ψ in a radial direction orthogonal to x :

$$\delta^{out}(x, y) = \inf\{\lambda > 0 | (x, \lambda^{-1}y) \in \Psi\} \equiv (\lambda(x, y))^{-1}. \quad (2.13)$$

For all $(x, y) \in \Psi$, $\delta^{out}(x, y) \leq 1$. If either $\delta^{in}(x, y) = 1$ or $\delta^{out}(x, y) = 1$ then (x, y) belongs to the frontier of Ψ and the firm is technically efficient.

As pointed out in Simar and Wilson (2001), no behavioral assumptions are necessary for measuring technical efficiency. From a purely technical viewpoint, either the input or the output distance function can be used to measure technical efficiency - the only difference is in the direction in which distance to the technology is measured. The way of looking at the frontier will typically depend on the context of the application. For instance, if the outputs are exogenous and not under the control of the Decision Makers (*e.g.* as in most of the public services), input efficiency will be of main interest, since the inputs are the only elements under the control of the managers. But even in this case, both measures are available.

2.4 A taxonomy of efficient frontier models

The analysis of the existent literature is a necessary step for the advancement of a discipline. This is particularly true for the field of efficiency and productivity research that in the last decades has known an exponential increasing in the number of methodological and applied works. For a DEA bibliography over 1978-1992, see Seiford (1994, 1996) and for an extension till 2001 see Gattoufi, Oral and Reisman (2004). In Cooper, Seiford and Tone (2000) about 1,500 DEA references are reported. Other bibliographic studies include: Emrouznejad (2001) and Taveres (2002).

As a consequence, a comprehensive review of the overall literature would require another whole work. Therefore, the aim of this section is to propose a general taxonomy of efficient frontier models that gives an overview on the different approaches presented in literature for estimating the efficient frontier of a production possibility set. Here the review could be biased toward the nonparametric approach, due to our commitment and involvement with nonparametric methods most. Anyway, we give several references also on the parametric approach that could be useful for those interested in it.

In the previous section we described the economic model underlying the frontier analysis framework based on the Activity Analysis Model. This model is based on some representations of the production set Ψ on which we can impose different axioms. Nevertheless, the production set Ψ , the boundary of the input requirement set $\partial C(y)$ and of the output correspondence set $\partial P(x)$, together with the efficiency scores in the input and output space, $\theta(x, y)$ and $\lambda(x, y)$, are unknown.

The econometric problem is thus how to estimate Ψ , and then $\partial C(y)$, $\partial P(x)$, $\theta(x, y)$, $\lambda(x, y)$, from a random sample of production units $\mathcal{X} = \{(X_i, Y_i) \mid i = 1, \dots, n\}$.

Starting from the first empirical application of Farrell (1957) several different approaches for efficient frontier estimation and efficiency score calculation have been developed.⁸

In Figure 2.1 we propose an outline of what we believe have been the most influential works in productivity and efficiency analysis, starting from the pioneering work by Farrell (1957). Of course, our outline is far from being complete and *all-inclusive*. Figure 2.1 shows some of the articles, books and special issues of journals (*i.e.* *Journal of Econometrics* JE, *Journal of Productivity Analysis* JPA, *European Journal of Operational Research*, EJOR) that have mainly influenced the writing of this work, trying to balance them according to the adopted approach.

As it is evident from Figure 2.1 we have taken into consideration mainly the nonparametric approach as we believe that thanks to its last developments, it can be considered as being very flexible and very useful for modeling purpose.

We may classify efficient frontier models according to the following *criteria*:⁹

- 1 The specification of the (functional) form for the *frontier function*;
- 2 The presence of noise in the sample data;
- 3 The type of data analyzed.

Based on the first *criterion* (functional form of the frontier) is the classification in:

- *Parametric Models*. In these models, the attainable set Ψ is defined through a *production frontier function*, $g(x, \beta)$, which is a known mathematical function depending on some k unknown parameters, *i.e.* $\beta \in \mathcal{R}^k$, where generally y is univariate, *i.e.* $y \in \mathcal{R}_+$. The main advantages of this approach are the economic interpretation of parameters and the statistical properties of estimators; more critical are the choice of the function $g(x, \beta)$ and the handling of multiple inputs, multiple outputs cases (for more on this latter aspect see Section 4.7 below where we introduce multivariate parametric approximations of nonparametric and robust frontiers).
- *Nonparametric Models*. These models do not assume any particular functional form for the frontier function $g(x)$. The main pros of this approach are the robustness to model choice and the easy handling of multiple inputs, multiple outputs case; their main limitations are the estimation of unknown functional and the *curse of dimensionality*¹⁰, typical of nonparametric methods.

⁸For an introduction see *e.g.*, Coelli, Rao and Battese (1998) and Thanassoulis (2001).

⁹These *criteria* follow Simar and Wilson (2006b), where a comprehensive statistical approach is described.

¹⁰The curse of dimensionality, shared by many nonparametric methods, means that to avoid large variances and wide confidence interval estimates a large quantity of data is needed.

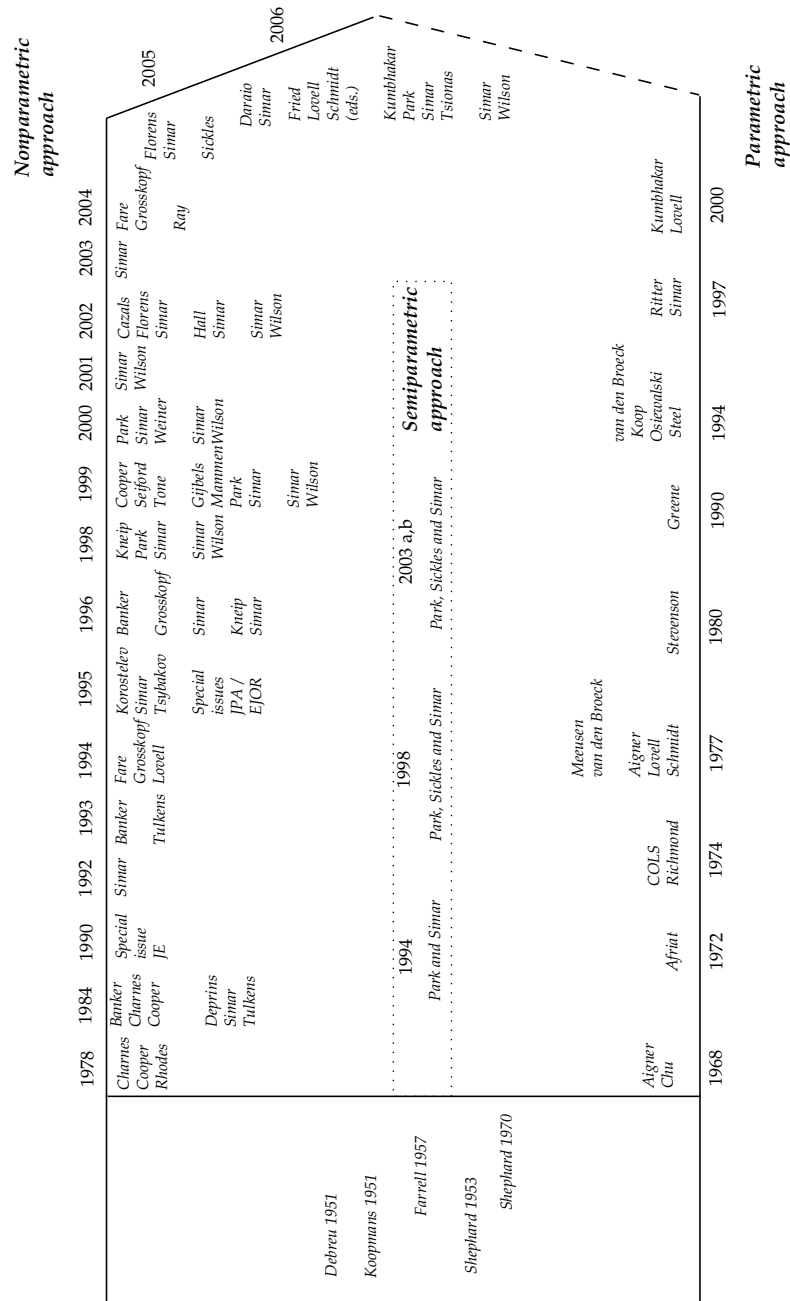


Figure 2.1. An overview of the literature on efficient frontier estimation.

Based on the second *criterium* (presence of noise) is the classification in:

- *Deterministic Models*, which assume that all observations (X_i, Y_i) belong to the production set, *i.e.*

$$Prob\{(X_i, Y_i) \in \Psi\} = 1$$

for all $i = 1, \dots, n$. The main weakness of this approach is the sensitivity to “super-efficient” outliers. Robust estimators are able to overcome this drawback.

- *Stochastic Models*, in which there might be noise in the data, *i.e.* some observations might lie outside Ψ . The main problem of this approach is the identification of noise from inefficiency.

Based on the third *criterium* (type of data analyzed) is the classification in:

- *Cross-sectional Models*, in which the data sample is done by observations on n firms or DMUs (Decision Making Units):

$$\mathcal{X} = \{(X_i, Y_i) | i = 1, \dots, n\}$$

- *Panel Data Models*, in which the observations on the n firms are available over T periods of time:

$$\mathcal{X} = \{(X_{it}, Y_{it}) \mid i = 1, \dots, n; t = 1, \dots, T\}.$$

Panel data allow the measurement of *productivity change* as well as the estimation of technical progress or regress.

Generally speaking, productivity change occurs when an index of outputs changes at a different rate than an index of inputs does. Productivity change can be calculated using index number techniques to construct a Fisher (1922) or Tornqvist (1936) productivity index. Both these indices require quantity and price information, as well as assumptions concerning the structure of technology and the behavior of producers. Productivity change can also be calculated using nonparametric techniques to construct a Malmquist (1953) productivity index. These latter techniques do not require price information or technological and behavioral assumptions, but they require the estimation of a representation of production technology. Nonparametric techniques are able not only to calculate productivity change, but also to identify the sources of measured productivity change.

A survey of the theoretical and empirical work on Malmquist productivity indices can be found in Färe, Grosskopf and Russell (1998). On the theoretical side the survey includes a number of issues that have arisen since the Malmquist productivity index was proposed by Caves, Christensen and Diewert

(1982). These issues include the definition of the Malmquist productivity index; although all are based on the distance functions that Malmquist employed to formulate his original quantity index, variations include the geometric mean form used by Färe, Grosskopf, Lindgren and Roos (1989) and the quantity index form by Diewert (1992). The survey of the empirical literature presents studies on the public sector, banking, agriculture, countries and international comparisons, electric utilities, transportation, and insurance. See also Lovell (2003), and Grosskopf (2003) for an historical perspective and an outline of the state of the art in this area.

Although productivity change is not the main focus of FDH, it can be inferred from information on efficiency change and technical change that is revealed by FDH. The technique was developed by Tulkens that named it “sequential FDH”. For an illustration of the sequential FDH see Lovell (1993, pp. 48-49). On this topic see also Tulkens and Vanden Eeckaut (1995a, 1995b).

By combining the three *criteria* mentioned above, several models have been studied in literature:

- *Parametric Deterministic Models*, see e.g. Aigner and Chu (1968), Afriat (1972), Richmond (1974), Schmidt (1976) and Greene (1980) for cross-sectional and panel data;
- *Parametric Stochastic Models*, most of these techniques are based on the maximum likelihood principle, following the pioneering works of Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977). For a recent review see Kumbhakar and Lovell (2000). In the context of panel data, stochastic models (see Schmidt and Sickles, 1984, and Cornwell, Schmidt, and Sickles, 1990) have *semiparametric* generalizations, in which a part of the model is parametric and the rest is nonparametric (see Park and Simar, 1994; Park, Sickles and Simar, 1998; and Park, Sickles and Simar, 2003a, b).
- *Nonparametric Deterministic Models* for cross-sectional and panel data. Traditional references on these models include: Färe, Grosskopf and Lovell (1985, 1994), Fried, Lovell and Schmidt (1993), and Charnes, Cooper, Lewin and Seiford, 1994. Recent and updated references are Cooper, Seiford and Tone (2000), Ray (2004) and Färe and Grosskopf (2004).
- *Nonparametric Stochastic Models* for cross-sectional data (see Hall and Simar, 2002; Simar, 2003b; Kumbhakar, Park, Simar and Tsionas, 2004) and panel data (see Kneip and Simar, 1996; and Henderson and Simar, 2005).

The mainly used approaches in empirical works are the nonparametric (deterministic) frontier approach and the (parametric) stochastic frontier approach.

In the following, when we refer to nonparametric frontier approach we indicate the deterministic version of it; when we talk about stochastic frontier approach we refer to its parametric version.

The nonparametric frontier approach, based on envelopment techniques (DEA FDH), has been extensively used for estimating efficiency of firms as it relies only on very few assumptions for Ψ . On the contrary, the stochastic frontier approach (SFA) allows the presence of noise but it demands parametric restrictions on the shape of the frontier and on the Data Generating Process (DGP) in order to permit the identification of noise from inefficiency and the estimation of the frontier. Fried, Lovell and Schmidt (2006) offer an updated presentation of both approaches. A statistical approach which unifies parametric and nonparametric approaches can be found in Simar and Wilson (2006b).

2.5 The nonparametric frontier approach

In this section we introduce the most known nonparametric estimators of efficient frontiers.

As we have seen in Section 2.3 devoted to the presentation of the economic model, we can equivalently look at the efficient boundary of Ψ from the input space or from the output space.

The *input oriented* framework, based on the input requirement set and its efficient boundary, aims at reducing the input amounts by as much as possible while keeping at least the present output levels. This is also called “input-saving” approach to stress the fact that the outputs level remains unchanged and input quantities are reduced proportionately till the frontier is reached. This is a framework generally adopted when the *decision maker* can control the inputs but has not the control of the outputs. For instance, this is the case of public enterprises which are committed to offer some public services and are interested in the management of the inputs, in the sense of their minimization.

Alternatively, we can take into account the output space and look at the output correspondence set and its efficient boundary. The *output oriented* framework looks at maximize output levels under at most the present input consumption. This approach is also known as “output-augmenting” approach, because it holds the input bundle unchanged and expand the output level till the frontier is reached. In practice, whether the input or output-oriented measure is more appropriate would depend on whether input conservation is more important than output augmentation.

For the relation existent among input and output efficiency measures, see Deprins and Simar (1983).

The main nonparametric estimators available are the Data Envelopment Analysis (DEA) and the Free Disposal Hull (FDH) which we describe in the subsections that follow.

2.5.1 Data Envelopment Analysis (DEA)

The DEA estimator of the production set, initiated by Farrell (1957) and operationalized as linear programming estimators by Charnes, Cooper and Rhodes (1978), assumes the free disposability and the convexity of the production set Ψ . It involves measurement of efficiency for a given unit (x, y) relative to the boundary of the convex hull of $\mathcal{X} = \{(X_i, Y_i), i = 1, \dots, n\}$:

$$\begin{aligned} \hat{\Psi}_{DEA} = & \left\{ (x, y) \in \mathcal{R}_+^{p+q} \mid y \leq \sum_{i=1}^n \gamma_i Y_i; x \geq \sum_{i=1}^n \gamma_i X_i, \text{ for } (\gamma_1, \dots, \gamma_n) \right. \\ & \left. \text{s.t. } \sum_{i=1}^n \gamma_i = 1; \gamma_i \geq 0, i = 1, \dots, n \right\} \end{aligned} \quad (2.14)$$

$\hat{\Psi}_{DEA}$ is thus the smallest free disposal convex set covering all the data.

The $\hat{\Psi}_{DEA}$ in (2.14) allows for Variable Returns to Scale (VRS) and is often referred as $\hat{\Psi}_{DEA-VRS}$ (see Banker, Charnes and Cooper, 1984). It may be adapted to other returns to scale situations. It allows for:

- *Constant Returns to Scale* (CRS) if the equality constrained $\sum_{i=1}^n \gamma_i = 1$ in (2.14) is dropped;
- *Non Increasing Returns to Scale* (NIRS) if the equality constrained $\sum_{i=1}^n \gamma_i = 1$ in (2.14) is changed in $\sum_{i=1}^n \gamma_i \leq 1$;
- *Non Decreasing Returns to Scale* (NDRS) if the equality constrained $\sum_{i=1}^n \gamma_i = 1$ in (2.14) is modified in $\sum_{i=1}^n \gamma_i \geq 1$.

The estimation of the input requirement set is given for all y by: $\hat{C}(y) = \{x \in \mathcal{R}_+^p \mid (x, y) \in \hat{\Psi}_{DEA}\}$ and $\partial \hat{C}(y)$ denotes the estimator of the input frontier boundary for y .

For a firm operating at level (x_0, y_0) the estimation of the input efficiency score $\theta(x_0, y_0)$ is obtained by solving the following linear program (here and hereafter we consider the VRS case):

$$\hat{\theta}_{DEA}(x_0, y_0) = \inf \left\{ \theta \mid (\theta x_0, y_0) \in \hat{\Psi}_{DEA} \right\} \quad (2.15)$$

$$\begin{aligned} \hat{\theta}_{DEA}(x_0, y_0) = \min \left\{ \theta \mid y_0 \leq \sum_{i=1}^n \gamma_i Y_i; \theta x_0 \geq \sum_{i=1}^n \gamma_i X_i; \theta > 0; \right. \\ \left. \sum_{i=1}^n \gamma_i = 1; \gamma_i \geq 0; i = 1, \dots, n \right\}. \end{aligned} \quad (2.16)$$

$\hat{\theta}(x_0, y_0)$ measures the radial distance between (x_0, y_0) and $(\hat{x}^\partial(x_0|y_0), y_0)$ where $\hat{x}^\partial(x_0|y_0)$ is the level of the inputs the unit should reach in order to

be on the “efficient boundary” of $\hat{\Psi}_{DEA}$ with the same level of output, y_0 , and the same proportion of inputs; *i.e.* moving from x_0 to $\hat{x}^\partial(x_0|y_0)$ along the ray θx_0 . The projection of x_0 on the efficient frontier is thus equal to $\hat{x}^\partial(x_0|y_0) = \hat{\theta}(x_0, y_0)x_0$.

For the output oriented case, the estimation is done, *mutatis mutandis*, following the previous steps. The output correspondence set is estimated by: $\hat{P}(x) = \{y \in \mathcal{R}_+^q | (x, y) \in \hat{\Psi}_{DEA}\}$ and $\partial\hat{P}(x)$ denotes the estimator of the output frontier boundary for x .

The estimator of the output efficiency score for a given (x_0, y_0) is obtained by solving the following linear program:

$$\hat{\lambda}_{DEA}(x_0, y_0) = \sup\{\lambda \mid (x_0, \lambda y_0) \in \hat{\Psi}_{DEA}\}, \quad (2.17)$$

$$\begin{aligned} \hat{\lambda}_{DEA}(x_0, y_0) = \max\left\{\lambda \mid \lambda y_0 \leq \sum_{i=1}^n \gamma_i Y_i; \quad x_0 \geq \sum_{i=1}^n \gamma_i X_i; \quad \lambda > 0; \right. \\ \left. \sum_{i=1}^n \gamma_i = 1; \quad \gamma_i \geq 0; \quad i = 1, \dots, n\right\}. \end{aligned} \quad (2.18)$$

In Figure 2.2 we display the DEA estimator and illustrate the concept of *slacks* through an example. If we look at the left panel assuming that all firms produce the same level of output, we can see that the DMU E could actually produce 1 unit of y with less input x_1 , *i.e.*, it could reduce x_1 by one unit (from 4 to 3) moving from E to D. This is referred to as *input slack*: although the DMU is technical efficient, there is a *surplus* of input x_1 .¹¹ In general, we say that there is slack in input j of DMU i , *i.e.*, x_i^j , if:

$$\sum_{i=1}^n \gamma_i x_i < x_i^j \hat{\theta}(x_i, y_i) \quad (2.19)$$

is true for some solution value of γ_i , $i = 1, \dots, n$ (see Färe, Grosskopf and Lovell, 1994, for more details).

The same kind of reasoning can be done for the output oriented case, *i.e.* the DMU L could increase the production of y_1 moving from L to M. See Figure 2.2, right panel for a graphical illustration.

Slacks may happen for DEA estimates (as shown in Figure 2.2), as well as for FDH estimates (presented in the next section). It is interesting to note that if the true production set Ψ has no slacks, than slacks are only a small sample problem. Nevertheless, it is always useful to report slacks whenever they are

¹¹Remember the “possibility of destroying goods without costs” underlying the frontier representation of the economic model.

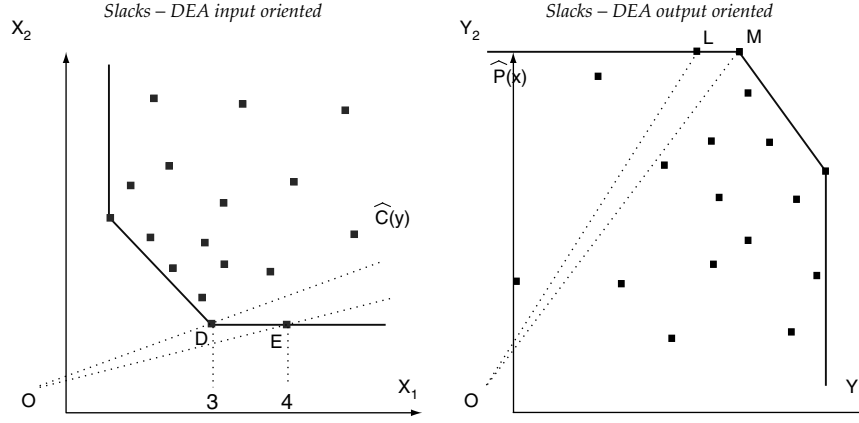


Figure 2.2. Input and Output slacks.

there. It is left to the analyst to decide if it is better to correct for the slacks or just point them.

Once the efficiency measures have been computed, several interesting analysis could be done, such as the inspection of the distribution of efficiency scores and the analysis of the “best performers” or efficient facet of the frontier closer to the analysed DMU, generally called *peer-analysis*, to study the technical efficient units and try to learn from them.

2.5.2 Free Disposal Hull (FDH)

The FDH estimator, proposed by Deprins, Simar and Tulkens (1984), is a more general version of the DEA estimator as it relies only on the free disposability assumption for Ψ , and hence does not restrict itself to convex technologies. This seems an attractive property of FDH since it is frequently difficult to find a good theoretical or empirical justification for postulating convex production sets in efficiency analysis. At this purpose, Farrell (1959) indicates indivisibility of inputs and outputs and economies of scale and specialization as possible violations of convexity. It is important to note also that if the *true* production set is convex then the DEA and FDH are both consistent estimators; however, as pointed later in this section, FDH shows a lower rate of convergence (due to the less assumptions it requires) with respect to DEA. On the contrary, if the *true* production set is not convex, then DEA is not a consistent estimator of the production set, while FDH is consistent.

The FDH estimator measures the efficiency for a given point (x_0, y_0) relative to the boundary of the Free Disposal Hull of the sample $\mathcal{X} = \{(X_i, Y_i), i = 1, \dots, n\}$. The Free Disposal Hull of the set of observations (*i.e.* the FDH

estimator of Ψ) is defined as:

$$\hat{\Psi}_{FDH} = \left\{ (x, y) \in \mathcal{R}_+^{p+q} \mid y \leq Y_i; x \geq X_i, (X_i, Y_i) \in \mathcal{X} \right\}. \quad (2.20)$$

It is the union of the all positive orthants in the inputs and of the negative orthants in the outputs whose origin coincides with the observed points $(X_i, Y_i) \in \mathcal{X}$ (Deprins, Simar and Tulkens, 1984). See Figures 2.3 and 2.4 where the FDH estimator is compared with the DEA estimator of the input and output requirement sets, respectively.

The efficiency estimators, in this framework, are obtained (as for the DEA case) using a “plug-in principle”, *i.e.*, by substituting the unknown quantities (in this case Ψ) by their estimated values (here $\hat{\Psi}_{FDH}$, for the DEA case $\hat{\Psi}_{DEA}$).

The estimated input requirement set and the output correspondence set are the following:

$$\hat{C}(y) = \{x \in \mathcal{R}_+^p \mid (x, y) \in \hat{\Psi}_{FDH}\},$$

$$\hat{P}(x) = \{y \in \mathcal{R}_+^q \mid (x, y) \in \hat{\Psi}_{FDH}\}.$$

Their respective efficient boundaries are:

$$\partial \hat{C}(y) = \{x \mid x \in \hat{C}(y), \theta x \notin \hat{C}(y) \forall 0 < \theta < 1\},$$

$$\partial \hat{P}(x) = \{y \mid y \in \hat{P}(x), \lambda y \notin \hat{P}(x) \forall \lambda > 1\}.$$

Hence, the estimated input efficiency score for a given point $(x_0, y_0) \in \Psi$ is:

$$\begin{aligned} \hat{\theta}_{FDH}(x_0, y_0) &= \inf \left\{ \theta \mid \theta x_0 \in \hat{C}(y_0) \right\} \\ &= \inf \left\{ \theta \mid (\theta x_0, y_0) \in \hat{\Psi}_{FDH} \right\}, \end{aligned} \quad (2.21)$$

and the estimated output efficiency score of (x_0, y_0) is given by:

$$\begin{aligned} \hat{\lambda}_{FDH}(x_0, y_0) &= \sup \left\{ \lambda \mid \lambda y_0 \in \hat{P}(x_0) \right\} \\ &= \sup \left\{ \lambda \mid (x_0, \lambda y_0) \in \hat{\Psi}_{FDH} \right\}. \end{aligned} \quad (2.22)$$

It is clear that for a particular point (x_0, y_0) , the estimated distance to the frontiers are evaluated by means of the distance, in the input space (“input oriented”) from this point to the estimated frontier of the input requirement set ($\partial \hat{C}(y)$), and in the output space (“output oriented”) by the distance from (x_0, y_0) to the estimated frontier of the output correspondence set ($\partial \hat{P}(x)$).

It is worthwhile to note that the FDH attainable set in (2.20) can also be characterized as the following set:

$$\hat{\Psi}_{FDH} = \left\{ (x, y) \in \mathcal{R}_+^{p+q} \mid y \leq \sum_{i=1}^n \gamma_i Y_i; x \geq \sum_{i=1}^n \gamma_i X_i, \sum_{i=1}^n \gamma_i = 1; \right. \\ \left. \gamma_i \in \{0, 1\}, i = 1, \dots, n \right\}. \quad (2.23)$$

Therefore the efficiencies can be estimated by solving the following integer linear programs; for the input-oriented case we have:

$$\hat{\theta}_{FDH}(x_0, y_0) = \min \left\{ \theta \mid y_0 \leq \sum_{i=1}^n \gamma_i Y_i; \theta x_0 \geq \sum_{i=1}^n \gamma_i X_i, \sum_{i=1}^n \gamma_i = 1; \right. \\ \left. \gamma_i \in \{0, 1\}, i = 1, \dots, n \right\}, \quad (2.24)$$

and for the output-oriented case:

$$\hat{\lambda}_{FDH}(x_0, y_0) = \max \left\{ \lambda \mid \lambda y_0 \leq \sum_{i=1}^n \gamma_i Y_i; x_0 \geq \sum_{i=1}^n \gamma_i X_i, \sum_{i=1}^n \gamma_i = 1; \right. \\ \left. \gamma_i \in \{0, 1\}, i = 1, \dots, n \right\}. \quad (2.25)$$

The latter expressions allow to make the comparison easier between the FDH and the DEA estimators (compare for instance (2.23) with (2.14)).

Figure 2.3 illustrates the estimation of the input requirement set $C(y)$ and of its boundary $\partial C(y)$ through FDH and DEA methods. The dashed line represents the FDH estimation of $\partial C(y)$, while the solid line shows the DEA estimation of it. The squares are the observations. The DEA and FDH estimates of efficiency score of production unit B, in Figure 2.3, are respectively: $\hat{\theta}_{DEA}(x_0, y_0) = |OB''|/|OB| \leq 1$, $\hat{\theta}_{FDH}(x_0, y_0) = |OB'|/|OB| \leq 1$.

In Figure 2.4 we show the FDH and DEA estimation of the output correspondence set $P(x)$ and its boundary $\partial P(x)$. The dash-dotted line represents the FDH estimator of $\partial P(x)$, while the solid line the DEA estimator of it. The black squares, as before, represent the DMUs. For firm B, the estimates of its efficiency score, in output oriented framework, are: $\hat{\lambda}_{FDH}(x_0, y_0) = |OB'|/|OB| \geq 1$, $\hat{\lambda}_{DEA}(x_0, y_0) = |OB''|/|OB| \geq 1$.

Practical computation of the FDH

In practice, the FDH estimator is computed by a simple vector comparison procedure that amounts to a complete enumeration algorithm proposed in Tulkens (1993), which is now explained.

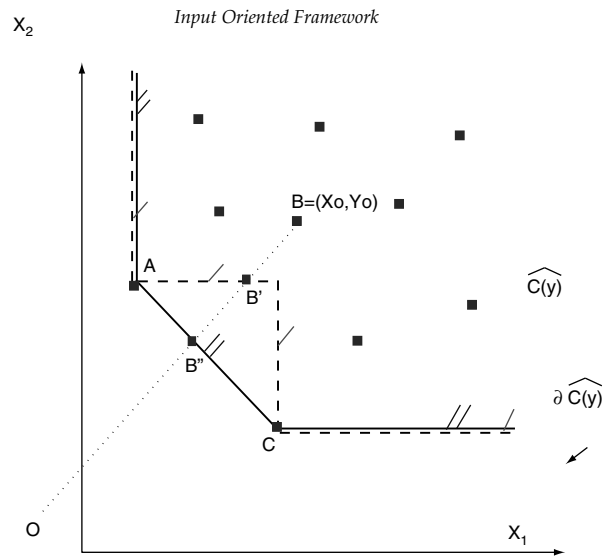


Figure 2.3. FDH and DEA estimation of $C(y)$ and $\partial C(y)$.

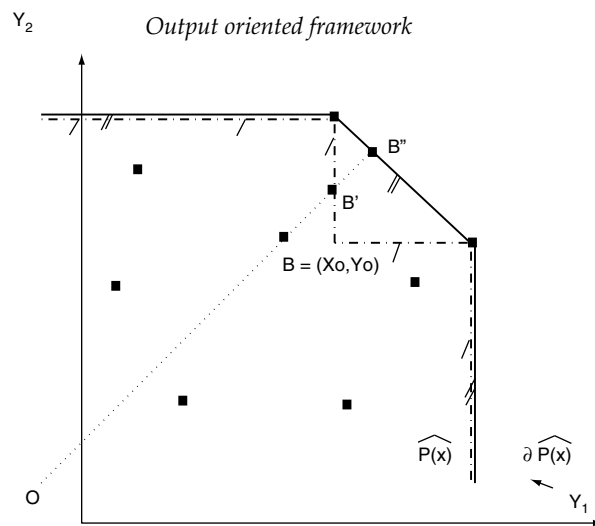


Figure 2.4. FDH and DEA estimation of $P(x)$ and $\partial P(x)$.

For a DMU (x_0, y_0) , in a first step, the set of observations which dominates it is determined, and then the estimate of its efficiency score, relative to the dominating facet of $\hat{\Psi}$ is computed. In the simplest case, with a technology characterized by one input and one output, the set of observations which dominate (x_0, y_0) is defined as:

$$D_0 = \left\{ i \mid (X_i, Y_i) \in \mathcal{X}, X_i \leq x_0, Y_i \geq y_0 \right\}. \quad (2.26)$$

The “input oriented” efficiency estimate is done through:

$$\hat{\theta}_{FDH}(x_0, y_0) = \min_{i \in D_0} \left(\frac{X_i}{x_0} \right), \quad (2.27)$$

and the “output oriented” efficiency is computed via:

$$\hat{\lambda}_{FDH}(x_0, y_0) = \max_{i \in D_0} \left(\frac{Y_i}{y_0} \right). \quad (2.28)$$

It has to be noted that as $X_i \leq x_0$ then $\hat{\theta}_{FDH} \leq 1$. As for the input-oriented case, from the fact that $Y_i \geq y_0$ follows that $\hat{\lambda}_{FDH} \geq 1$.

In a multivariate setting, the expression (2.21) can be computed through:

$$\hat{\theta}_{FDH}(x_0, y_0) = \min_{i \in D_0} \left\{ \max_{j=1, \dots, p} \left(\frac{X^{i,j}}{x_0^j} \right) \right\}, \quad (2.29)$$

where $X^{i,j}$ is the j^{th} component of $X^i \in \mathcal{R}_+^p$ and x_0^j is the j^{th} component of $x_0 \in \mathcal{R}_+^p$.

It is a *maximin* procedure (for the “input oriented” framework): the “max” part of the algorithm identifies the most dominant DMUs relative to which a given DMU is evaluated. Once the most dominant DMUs are identified, slacks are calculated from the “min” part of the algorithm.

The multivariate computation of expression (2.22) is done by:

$$\hat{\lambda}_{FDH}(x_0, y_0) = \max_{i \in D_0} \left\{ \min_{j=1, \dots, q} \left(\frac{Y^{i,j}}{y_0^j} \right) \right\} \quad (2.30)$$

where $Y^{i,j}$ is the j^{th} component of $Y^i \in \mathcal{R}_+^q$ and y_0^j is the j^{th} component of $y_0 \in \mathcal{R}_+^q$.

The FDH estimator has been applied in several contexts. For a detailed presentation of FDH concepts see Vanden Eeckaut (1997).

Recently, some authors have raised explicit doubts about the economic meaning of FDH, but from the exchange between Thrall (1999) and Cherchye, Kuosmanen and Post (2000), published on the *Journal of Productivity Analysis*, it

emerged that FDH can be economically more meaningful than convex monotone hull, also under non-trivial alternative economic conditions.

Hence, FDH technical efficiency measures remain meaningful for theories of the firm that do allow for imperfect competition or uncertainty (see *e.g.* Kuosmanen and Post, 2001, and Cherchye, Kuosmanen and Post, 2001).

One of the main drawbacks of deterministic frontier models (DEA /FDH based) is the influence of “super-efficient” outliers.

This is a consequence of the fact that the efficient frontier is determined by sample observations which are extreme points. Simar (1996) points out the need for identifying and eliminating outliers when using deterministic models. If they cannot be identified, the use of stochastic frontier models is recommended.

See Figure 2.5 for an illustration of the influence of outliers in case of FDH estimation. The same is valid for the DEA case. If point A is an extreme point, outlying the cloud of other points, the estimated efficient frontier is strongly influenced by it. In fact, in Figure 2.5, the solid line is the frontier that envelops point A, while the dash-dotted line does not envelop point A.

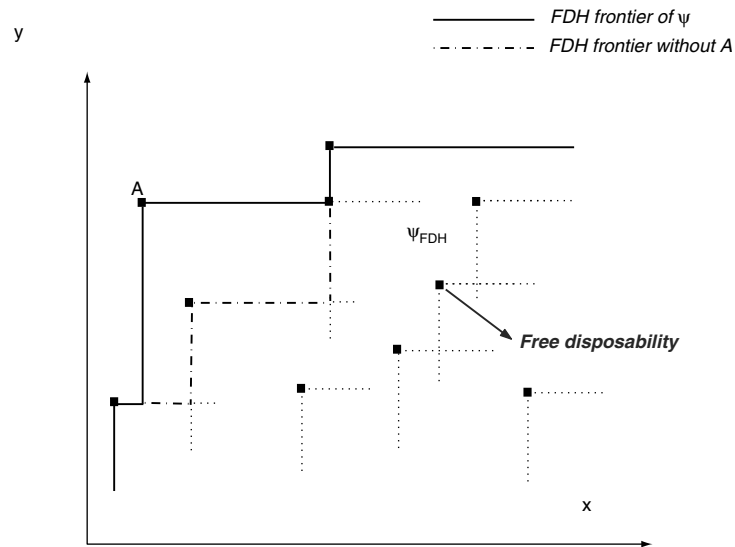


Figure 2.5. Influence of outliers on the FDH estimation of the production set Ψ .

We will come back on this problem in Chapter 4 where we propose robust nonparametric approaches based on various nonparametric measures less influenced by extreme values and outliers, which have also nice statistical properties.

2.6 Recent developments in nonparametric efficiency analysis

In the following, we recall briefly some stream of works that have contributed to the latest advancement of the nonparametric efficiency literature.¹²

Sensitivity of results to data variation and discrimination (in a DEA framework)

The focus of studies on sensitivity and stability is the reliability of classification of DMUs into efficient and inefficient performers. Most analytical methods for studying the sensitivity of results to variations in data have been developed in a DEA framework.

After a first stream of works concentrated on developing solution methods and algorithms for conducting sensitivity analysis in linear programming, a second current of studies analysed data variations in only one input or one output for one unit at a time. A recent stream of works makes it possible to determine ranges within which all data may be varied for any unit before a reclassification from efficient to inefficient status (or *vice versa*) occurs, and for determining ranges of data variation that can be allowed when all data are varied simultaneously for all DMUs. For a review and some references see Cooper, Li, Seiford, Tone, Thrall and Zhu (2001).

As we have seen above, DEA models have a deterministic nature, meaning that they do not account for statistical noise. Some authors (*e.g.*, Land, Lovell and Thore, 1993; Olesen and Petersen, 1995) have proposed the application of the *chance-constrained programming* to the DEA problem in order to overcome its deterministic nature. The basic idea is that of make DEA stochastic by introducing a chance that the constraints on either the envelopment problem or the multiplier problem may be violated with some probability. However, the chance-constrained efficiency measurement requires a large amount of data in addition to inputs and outputs. Moreover, it is based on a strong distributional assumption on the process determining the chance of a constrained to be violated. The analyst in fact has to provide also information on expected values of all variables for all DMUs, and variance-covariance matrices for each variable across all DMUs. An alternative to this approach is given by a fuzzy programming approach to DEA and FDH efficiency measurement.

There is an increasing number of studies that apply the fuzzy set theory in productivity and efficiency contexts. In some production studies, the data that describe the production process cannot be collected accurately due to the fact that measurement systems have not been originally designated for the pur-

¹²See also Lovell (2001) and Fried, Lovell and Schmidt (2006) for a presentation of some recent fruitful research areas introduced in parametric and nonparametric approaches to efficiency analysis.

pose of collecting data and information that are useful for production studies. Sengupta (1992) was the first to introduce a fuzzy mathematical programming approach where the constraints and objective function are not satisfied crisply. Seaver and Triantis (1992) proposed a fuzzy clustering approach for identify unusual or extreme efficient behavior. Girod and Triantis (1999) implemented a fuzzy linear programming approach, whilst Triantis and Girod (1998), and Kao and Liu (1999) used fuzzy set theory, to let the traditional DEA and FDH account for inaccuracies associated with the production plans. A fuzzy pairwise dominance approach can be found in Triantis and Vanden Eeckaut (2000) where, a classification scheme that explicitly accounts for the degree of fuzziness (plausibility) of dominating units is reported.

According to a classification proposed by Angulo-Meza and Pereira Estellita Lins (2002), the methods for increasing discrimination within efficient DMUs in a DEA setting can be classified into two groups:

Methods with a priori information. In these methods, the information provided by a decision-maker or an expert about the importance of the variables can be introduced into the DEA models. There are three main methods devoted to incorporating a priori information or value judgments in DEA:

- *Weight restrictions.* The main objective of the weight restrictions methods is to establish bounds within which the weights can vary, preserving some flexibility/ uncertainty about the real value of the weights.¹³
- *Preference structure models.* These models have been introduced by Zhu (1996) within a framework of non-radial efficiency measures. In this approach, the target for inefficient DMUs is given by a preference structure (represented through some weights) expressed by the decision-maker.
- *Value efficiency analysis.* This method, introduced by Halme, Joro, Korhonen, Salo and Wallenius (2000), aims at incorporate the decision-maker's value judgements and preferences into the analysis, using a two stage procedure. The first stage identifies the decision maker's most preferred solutions through a multiple objective model. The second stage consists in the determination of the frontier based on the most preferred solutions chosen.

Methods that do not require a priori information. These family of models aims at increase discrimination in DEA without the subjectivity, the possibility of biased or wrong judgements, typical of the methods that introduce a priori

¹³See Allen, Athanassopoulos, Dyson and Thanassoulis (1997), and Pedraja-Chaparro, Salinas-Jimenes, Smith and Smith (1997) for a review of some methods within this approach, including direct weight restrictions, cone ratio models, assurance region and virtual inputs and outputs restrictions.

information. The main methods that minimize the intervention of the experts are:

- *Super efficiency*. Andersen and Petersen (1993) proposed this method to rank efficient DMUs.
- *Cross-evaluation*. The main idea of this method is to use DEA in a “peer-evaluation” instead of a classical “self evaluation” evaluated by the classical DEA models.
- *Multiple objective approach*. A Multiple Criteria Data Envelopment Analysis has been proposed by Li and Reeves (1999) to solve the problems of lack of discrimination and inappropriate weighting schemes in traditional DEA.

Extensions to the basic DEA Models

Directional distance functions have been introduced by Chambers, Chung and Färe, (1996) and are based on Luenberger (1992) benefit functions. These functions represent a kind of generalization of the traditional distance functions. Their application leads to measures of technical efficiency from the potential for increasing outputs while reducing inputs at the same time. In order to provide a measure of “directional” efficiency, a *direction*, along which the observed DMU is projected onto the efficient frontier of the production set, has to be chosen. This choice is arbitrary and of course affects the resulting efficiency measures. In addition, those measures are no more scale-invariant. See Färe and Grosskopf (2004) for more details on these “new directions” in efficiency analysis.

Examples of the literature that try to link DEA with a theoretical foundation or that try to overcome and generalize the economic assumptions underlying DEA include: Bogetoft (2000) which links the theoretically oriented agency, incentives and contracts literature with the more practical oriented efficiency measurement literature; and Briec, Kerstens and Vanden Eeckaut (2004a, b) which extend the duality properties to non-convex technologies and propose congestion-based measures in this framework.

Producers face uncertainty about technology reliability and performance. The structure of technology and the existence and magnitude of inefficiency are sensitive to the treatment of risk and uncertainty. On productivity measurement under uncertainty see Chambers and Quiggin (2000) and Chambers (2004).

Statistical inference in efficiency analysis

All what we have seen in the previous description of recent developments does not allow for a *statistical* sensitivity analysis, neither for rigorous statistical testing procedures. This is because the previous literature does not relies on a

statistical model; there is not, in fact, a definition of the Data Generating Process (DGP) and there is no room for statistical inference based on the construction of confidence intervals, estimation of the bias, statistical tests of hypothesis and so on.

There is instead a new approach, recently developed, which aims exactly at the analysis of the statistical properties of the nonparametric estimators, trying to overcome most limitations of traditional nonparametric methods and allowing for statistical inference and rigorous testing procedures. This literature is the main focus of this book. To the review of the statistical properties of nonparametric frontier estimators we devote the following Chapter 3. Chapter 4 deals in detail with a family of robust nonparametric measures of efficiency, which are more resistant to the influence of outliers and errors in data while having good statistical properties which let inference feasible in this complex framework. Finally, Chapter 5 illustrates and develop further the topic of conditional and robust measures of efficiency and an alternative way to evaluate the impact of external-environmental variables based on conditional measures of efficiency.



<http://www.springer.com/978-0-387-35155-1>

Advanced Robust and Nonparametric Methods in
Efficiency Analysis

Methodology and Applications

Daraio, C.; Simar, L.

2007, XXII, 248 p., Hardcover

ISBN: 978-0-387-35155-1