ON THE REGULATORY APPLICATION OF EFFICIENCY MEASURES[†]

Martín A. ROSSI
Centro de Estudios Económicos de la Regulación (CEER)
Departamento de Economía y Finanzas, UADE

Christian A. RUZZIER
Centro de Estudios Económicos de la Regulación (CEER)
Departamento de Economía y Finanzas, UADE

Abstract
The last decade has witnessed a change to more powerful incentive schemes and the adoption by a large number of regulators of some form of price cap regimes. The efficiency frontiers literature tackles the problem of measuring the X factor in a price cap regime with an RPI – X rule. However, that literature has by large focused solely on the theoretical aspects involved in the estimation of an efficient frontier. A thorough discussion of the empirical application of the theoretical concepts is, in a sense, missing. In this paper we address this issue and try to elaborate upon the applied aspects of efficiency measurement.

Resumen
La última década ha presenciado un cambio hacia esquemas más poderosos de incentivos, y la adopción por parte de gran número de reguladores de una regulación por precios máximos. La literatura de fronteras de eficiencia trata el problema de la medición del factor X en un régimen de precios máximos con una regla de RPI – X. Sin embargo, esta literatura se ha centrado en los aspectos teóricos de la estimación de una frontera. En cierto sentido, se carece de una discusión profunda sobre la aplicación de aquellos conceptos teóricos. En este trabajo encaramos esta discusión e intentamos profundizar en los aspectos empíricos de la medición de la eficiencia.

Códigos JEL: L5, L9.

I. Introduction

For decades, rate-of-return regulation has been the dominant practice in the regulation of utilities. This method, although allowing the firm to recover its costs and resulting in a lower cost of capital (due to the lower risk borne by the firms), provided little incentives for cost minimization among regulated firms. The last decade has witnessed a change to more powerful incentive schemes and the adoption by a large number of regulators of some form of price cap regimes.[1] The main purposes of a switch from rate-of-return regulation to price cap regulation have been to increase the incentives for firms to minimize their costs, and to ensure that, eventually, users benefit from these cost reductions –typically within 3-5 years after a regulatory price review. This objective requires the measurement of the expected efficiency gains that would lead to cost reductions at the firm level. The renewed attention given to productive efficiency is one of the main reasons for the increase in efforts to

[†] The authors wish to thank Martín Rodriguez Pardina, Antonio Estache, Carlos Romero, Iván Canay, and numerous participants at various seminars for very useful comments and helpful discussions. Any mistake is ours and, besides, this paper does not engage in any way the institution we are affiliated with.

measure efficiency in regulated sectors. Efficiency measures are no longer a side show as they were under rate-of-return regulation.

Efficiency gains of a firm can come from two main sources: shifts in the frontier reflecting efficiency gains at the sectoral level, and efficiency gains at the firm level reflecting a catching up effect. The latter are the gains to be made by firms not yet on the frontier. These firms should be able to achieve not only the industry gain (the shift of the frontier) but also specific gains offsetting firm specific inefficiencies. A regulator should bear in mind this decomposition when carrying out an efficiency analysis.

The efficiency frontiers literature tackles the problem of measuring both components of the X factor in a price cap regime with an RPI – X rule. However, that literature has by large focused solely on the theoretical aspects involved in the estimation of an efficient frontier. A thorough discussion of the empirical application of the theoretical concepts (which is the main interest of regulators) is, in a sense, missing. In this paper we address this issue and try to elaborate upon the applied aspects of efficiency measurement in a regulatory context.

The paper outline is as follows. Section II deals with the choices faced by a regulator willing to evaluate regulated firms' performances. Section III presents the consistency conditions that should be met by the efficiency measures to be useful to regulators, and discusses how to apply them in a regulatory setting. Finally, in Section IV, conclusions to this work are made.

II. Regulatory choices

An efficiency measure is, broadly speaking, the distance of the observed practice to the efficient frontier. The regulatory task of measuring efficiency would be greatly simplified *if* this frontier were known. Unfortunately, the regulator has no knowledge of the efficient frontier and thus has to estimate it. This should constitute the main concern of the regulator when attempting to measure the efficiency of regulated firms, for different estimates of the frontier would lead to potentially distinct assessments (as would different distance concepts).

There are a number of choices a regulator has to make in order to be able to estimate an efficient frontier, and the options she makes will potentially give rise to different performance evaluations. It is important that the regulator can count on a sound set of arguments in favor of the choices made. The main goal of this study is to provide with the empirically relevant arguments that support each decision.

The first decision is how to construct the efficient frontier. There are basically two alternatives: (i) a theoretically defined function based on engineering knowledge of the process of the industry, or (ii) an empirical function constructed on estimates based on observed data. Next comes a decision about the relevant efficiency concept to be measured: (i) productive (or overall), (ii) technical[2], or (iii) allocative. A choice related to the previous one has to do with the kind of relationship that is going to be estimated: (i) a cost function (productive efficiency estimates), or (ii) a production function (only technical efficiency measures).

There still remain other choices to be made. Is the frontier going to be estimated with parametric or non-parametric techniques? Is the distance to the frontier going to be attributed to inefficiency, or to random noise, or to some combination of both?

Having solved all the questions regarding the methodology to be employed, the regulator still has to decide upon the variables that should be included in the analysis. Which are the outputs of the industry? Which are the inputs? Are there variables beyond the firms' control?

The efficiency literature has dealt with these questions in depth, although in too much a theoretical way. Regulatory application has not been such an important issue. In the remainder of this section we discuss the pros and cons of all the preceding alternatives, not only from a theoretical point of view, but also from an empirical regulatory standpoint.

II.1. Theoretical function or best-observed practice

Modern regulatory regimes are focused on improving efficiency through incentive mechanisms. Among these, yardstick competition is a must. Yardstick competition, originally proposed by Shleifer (1985), requires the horizontal separation of some of the stages of a natural monopoly in order to obtain comparative information on relative efficiency levels of the firms. This information can then be used to set up tariffs for the regulated companies, allowing some efficiency gains to be passed on to consumers and preserving at the same time incentives for the firms to reduce their own costs. In other words, the regulator acting as the principal prefers to have several agents in order to reduce the existing asymmetry of information. In exchange for this superior knowledge some economies of scale and of scope are lost when the activity is separated into different units. If the firms were to be compared to a theoretically defined yardstick, however, the regulator would still be bearing the costs of lost scale economies, but it would not receive the benefits of increased information. In such a case, it would be better to compare the original natural monopoly (not divided) to that yardstick. Therefore, in those processes involving horizontal break-up of a natural monopoly, the best observed practice seems the natural choice.

Farrell (1957), in his path-breaking paper, argues in favor of using the best-observed practice:

> "In a first place, it is very difficult to specify a theoretical efficient function [...]. Thus, the more complex the process, the less accurate is the theoretical function likely to be. Also, partly because of this, and partly because the more complex the process, the more scope it allows to human frailty, the theoretical function is likely to be wildly optimistic. If the measures are to be used as some sort of yardstick for judging the success of individuals plants, firms, or industries, this is likely to have unfortunate psychological effects; it is far better to compare performances with the best actually achieved than with some unattainable level" (Farrell, 1957, p. 255).

In accordance to Farrell's suggestion, the growing practice for regulatory purposes is to analyze individual performances in relation with best-observed practice. This, for example, is the approach used in UK for regulating the water utilities, in Costa Rica for setting transport tariffs, and in Hungary to regulate telecommunications companies. Furthermore, in Norway, where there are sixty transmission and two hundred distribution utilities, regulators have taken affirmative steps to employ this approach in setting rates. The Norwegian Resources and Energy Administration has devised a software that it distributes to the regulated utilities so that they can perform their own efficiency analysis, which is then used in setting utilities rates (Reiter et al., 1999). However, there are exceptions. In Chile (water sector), Peru (electricity) and Spain (electricity), for example, the frontier is calculated on the basis of engineering data instead of relying on best practice.[3]

The regulator should bear in mind that if efficiency is measured against best observed practice the result would be a measure of *relative* efficiency, where the firm is being compared with the other firms in the sample. Therefore, being found 100% efficient does not imply that a firm cannot enhance its performance; it just means that no other firm in the sample in performing as well as it is.

II.2. Cost versus production functions[4]

Productive or overall efficiency is the firm's ability to produce an output at minimum cost. To achieve that minimum cost the firm must produce the maximum output given its inputs (technical efficiency) and choose the appropriate input mix given the relative price of its inputs (allocative efficiency). Thus, productive efficiency requires both technical and allocative efficiency. Therefore, productive inefficiency will tend to be higher than technical inefficiency.

Related to the decision of what kind of efficiency concept is going to be used is the type of relation that is going to be estimated: a production function or a cost function. A production function displays the produced quantities as a function of the inputs employed and gives information on technical efficiency only, whereas a cost function shows the total cost of production as a function of the level of output/s and the input prices and allows for the estimation of the overall productive efficiency.[5] Whereas technical efficiency is a purely physical notion that can be measured without having to impose a behavioral objective on producers, cost or overall efficiency is an economic concept whose measurement requires the imposition of an appropriate behavioral objective (Kumbhakar and Lovell, 2000).

Up to this point we have discussed the main theoretical considerations. However, several other considerations need to be made to arrive at an application that is both feasible and reasonable, and not a mere artificial construct.

When choosing between the estimation of a production function or a cost function, it is important to bear in mind the peculiarities of the sector one is studying. An important feature of the regulated utilities is that, in general, the firms are under obligation to provide the service at the specified tariffs. Therefore, the firms must meet the demand for their service, and are not able to choose the level of output they will offer. Given the exogeneity of the output levels, the firm maximizes profit simply by minimizing the cost of producing a given level of output. Under this argument, a cost function specification is the correct one.[6]

However, cost function estimation has some drawbacks. Among these is the difficulty to obtain accurate information on input prices. Moreover, the estimation of cost frontiers involves the utilization of variables measured in monetary units, which could be a serious problem if one wishes to make international comparisons. Production functions, instead, only require variables measured in physical units (i.e. homogeneous among countries –or at least much more homogeneous). As a theoretical argument, one could add that whenever there is public ownership, the firms, in general, will not seek profit maximization as their main goal. As Pestieu and Tulkens (1990) argue, public enterprises do not share the same objectives and constraints that their private counterparts do, so their relative performance should only be compared on the basis of technical efficiency (a common ground).[7]

Although a model of economic efficiency (basically a cost minimization problem) needs price data (Lovell, 1993), a way around the problem of unavailable input prices is that of *cost efficiency* measurement, where a simple single input-multiple output model is built, using a measure of costs as the single input. The model yields as a result the proportion in which costs could be reduced, without changing the level of output/s.[8] Such an approach has been applied, for example, by Vanden Eeckaut et al. (1993) and Ramos and Sousa (1998) to municipalities in Belgium and Brazil, respectively; and by DTe (2000) to the Dutch network and supply businesses in the electricity sector.

If a cost function approach were chosen, it would still be necessary to define what kind of cost is going to be measured. If substitution possibilities exist whereby capital may be substituted for other inputs (and vice versa) one may find that companies with higher capital costs may have lower operating costs whereas companies with lower capital inputs may have higher operating costs. That is, regulating the industry using a restricted definition of costs (e.g. operating expenditures, OPEX, or capital expenditures, CAPEX) could lead to an inefficient allocation of resources.

Moreover, if OPEX is the chosen cost concept, care has to be given to the fact that different companies can pursue different accounting rules, and therefore include some costs items in OPEX that other firms do not consider, and this may bias the results (as DTe (2000) found). In such circumstances to judge comparative efficiency solely on the grounds of operating costs may give a misleading picture of the overall efficiency of companies with respect to the use of all inputs (Bosworth, Stoneman and Thanassoulis, 1996).

On the other hand, working with total costs in a panel data setting can have some problems if firms are not observed in a sufficiently large number of years, covering at least one investment cycle. In such a case, it could prove better to work solely with OPEX, for these would make a more homogeneous cost measure by avoiding the differences in total costs provoked by irregular investment outlays (CAPEX) of the firms in such a short period.[9]

II.3. Parametric versus non-parametric frontiers

Another decision regarding estimation refers to whether the frontier is assumed parametric or non-parametric. Parametric methods impose an a priori functional form to the frontier, whereas non-parametric methods do not. Parametric methods estimate a production or cost function by means of econometric tools. The most used non-parametric approach is the so-called Data Envelopment Analysis (DEA), which involves the use of linear programming techniques. In this methodology firms are considered efficient if there are no other firms, or linear combination of firms, which produce more of at least one output (given the inputs) or use less of at least one input (given the outputs). The DEA methodology, introduced by Charnes, Cooper and Rhodes (1978), seeks to determine which units (firms) form an envelopment surface or efficient frontier. The firms that lie on (determine) the surface are considered efficient, whereas the firms below the surface are termed inefficient, and their distance to the frontier provides a measure of their relative (in)efficiency.

There exist basically two types of envelopment surfaces (Ali and Seiford, 1993), the so-called constant returns to scale surface (CRS) and variable returns to scale surface (VRS). Their names indicate that an assumption about the type of returns to scale is associated to the choice of either surface. The efficient frontier thus constructed will be different according to the returns to scale assumption adopted.

One assumption involved in traditional DEA estimations is convexity of the set of feasible input-output combinations (Lovell, 1993). If this assumption is not robust, however, another methodology could be used: FDH (“free disposal hull”). FDH envelops the data more tightly and has a more restrictive notion of domination than DEA: a firm is dominated in FDH by a single observed efficient firm, whereas it is dominated in DEA by a hypothetical firm obtained from a linear (or convex) combination of a set of efficient producers. An advantage of FDH is that in practice no frontier needs to be computed. A potentially serious flaw of the methodology is that of “efficiency by default”, i.e., a potentially large number of firms could be declared efficient not because they are efficient, but because of the absence of firms with which the dominance comparisons need to be made.

Though this problem is reduced in DEA calculations, it does not disappear. An aspect worth noting is that the efficiency measures obtained with DEA can be very sensitive to the number of variables included in the model. As the ratio *number of variables/sample size* grows, the ability of DEA to discriminate among firms is sharply reduced, because it becomes more likely that a certain firm will find some set of weights to apply to its outputs and inputs which will make it appear as efficient (Yunos and Hawdon, 1997). That is to say, a lot of firms might be labeled 100% efficient not because they dominate other firms, but just because there are no other firms or combinations of firms against which they can be compared when there are so many dimensions.[10] This problem seems to be important in applied research: Rodríguez Pardina, Rossi and Ruzzier (1999), for instance, estimate a production function for a cross section of 53 firms in the electricity distribution sector in South America, finding that in the variable returns to scale model more than half of the utilities were 100% efficient, even though the model has only five variables.

DEA technical efficiency models can be oriented (i) to the proportional reduction of inputs –input orientation- or (ii) to the proportional augmentation of outputs –output orientation-, or they can be not oriented (in which the input reduction and the output augmentation needed to place a firm on the frontier are calculated). It is important to notice that –once a type of surface is chosen- the form of the efficient frontier will not change *whichever the orientation* selected; i.e. every orientation will identify the same firms as being efficient or inefficient. The differences between orientation will be seen in the efficiency scores, for each differently oriented model uses a different distance concept.

The choice between different orientations will depend on the particular features of the sector under study; e.g. if output is exogenous, considering output-oriented or not oriented models would be nonsense, for no increase in outputs can be achieved. In those circumstances, only an input orientation would be meaningful.[11]

The principal advantage of non-parametric approaches is that no functional form of the frontier is imposed a priori on the data. A drawback is that only a subset of the available data defines the efficient frontier, while the rest of the observations have no impact on the shape of the envelopment surface. Furthermore, non-parametric methods estimate the efficient frontier without making any assumption about the distribution of the error term. The estimations, therefore, lack statistical properties, thus rendering impossible the hypothesis testing. The parametric models, in turn, although allowing for hypothesis testing, might label inefficiency something that actually is a mispecification of the model. In order to account for this problem it is preferable to estimate a flexible function, like the translog, which is in fact a second order approximation to any arbitrary functional form.[12]

Parametric frontiers can be estimated by some variant of Ordinary Least Squares (OLS) or by Maximum Likelihood (ML). OLS estimates an average function whose constant term is then corrected to transform the estimated function into a frontier. Therefore, the estimation of the technological parameters gives equal weights to both efficient and inefficient firms. ML, on the other hand, incorporates a priori information on the distribution asymmetry of the error term, hence giving more importance to the efficient firms in the estimation of the slope parameters.

Bardhan, Cooper and Kumbhakar (1998), in a simulation study using Monte-Carlo methods, found that parametric methodologies yield estimates of the technological parameters that are significantly different from the true parameters, and attribute this outcome to the mixture of efficient and inefficient observations. To overcome this inconvenient, Arnold, Bardhan, Cooper and Kumbhakar (1996) suggest a joint use of parametric and non-parametric techniques in a two-stage procedure: a first stage involves the utilization of DEA to identify efficient and inefficient firms, and in a second stage, the firms identified as efficient are incorporated as dummy variables in a regression. Bardhan, Cooper and Kumbhakar (1998) found that this two-stage DEA-dummy variable approach yielded estimates that did not differ significantly from the true parameter values.

II.4. Deterministic versus stochastic approaches

Once decided upon the kind of frontier to be estimated (cost or production) and the estimation technique (mathematical programming or econometrics), the next step is to determine whether such frontier is to be considered deterministic or stochastic.[13] If the activity frontier is deterministic all the firms share the same frontier and every discrepancy between the individual firm performance and the frontier is considered due to inefficiency, thus completely ignoring the possibility of a single firm performance being affected not only by inefficiencies in the management of its resources but also by factors absolutely beyond its control and not considered as regressors. Besides, deterministic approaches are very sensitive to the presence of outliers. A single outlier (due for example to measurement errors) can have deep effects on the estimations, and this outlier problem cannot be solved just by increasing the sample size. Though this problem will be present both in DEA and parametric estimations, the effect is quite different in both approaches: in DEA estimates, the outlier problem has the effect of changing the entire frontier, shifting both the technology parameters and the efficiency measures. In the parametric approaches, though having a similar effect on the efficiency measures, the outlier problem has almost no effect on the slope parameters since they are estimated using information on all firms, not just the ones on the frontier.

Estimation of deterministic frontiers involves the utilization of a one-sided error term, which implies that it is possible to define accurately the minimum necessary cost to achieve a given level of output. Therefore, the actual cost is simply the least cost plus an inefficiency term (bound to be equal to or greater than zero by definition).[14]

It is worthwhile noting that the deterministic techniques are in a sense polar opposites of Ordinary Least Squares (OLS) estimates: OLS attributes all variation in output not associated to variations in inputs (production approach) to random shocks, whereas the deterministic approaches attribute all variation in output not associated to variations in inputs to technical inefficiency. An alternative to these polar cases would be a model that attributes variation in output not associated to variations in inputs to some combination of random shocks and technical inefficiency.

Following this idea, the works of Aigner, Lovell and Schmidt (1977) and Meeusen and van de Broeck (1977) came into the scene, proposing the so-called stochastic frontiers, which are based on the idea that the deviations from the frontier could be partially out of the control of the analyzed firm. This approach uses a mix of one-sided and two-sided errors; i.e., given an output level, there exists a minimum feasible cost, but this minimum is stochastic and not precise. The idea is that the external events which influence the cost function are normally distributed (the firm being faced to favorable or unfavorable conditions with given likelihood) instead of being constant. Once considered the likelihood of statistical noise, what remains is termed inefficiency.

It is worthwhile noting that this decomposition between statistical noise and inefficiency is precisely the nature of the moral hazard problem faced by an imperfectly informed regulator. That is, the regulator must establish which fraction of the observed differences between the firms' costs is due to inefficiency and which to external factors over which the firms have no control. The probability that some inefficiencies are erroneously classified as statistical noise is an important drawback in the regulatory context (Pollitt, 1995).

In the stochastic vs. deterministic dilemma OFWAT (the water and sewerage regulator in UK) explored both possibilities[15] in a number of research papers published for the 1994 Periodic Review and concluded that the deterministic approach was the most appropriate. This is because stochastic frontier models rely on too large a number of assumptions, which may not hold for the information collected from the companies. The deterministic approach does not require such strong assumptions (OFWAT, 1998).[16]

II.5. Panel data models

In general the stochastic frontier models with cross-sectional data are exposed to three serious drawbacks (Schmidt and Sickles, 1984). Firstly, the inefficiency term estimations, although unbiased, are not consistent, which really poses a problem if one bears in mind that the goal of the regulator is the estimation of the sample firms' inefficiency. Secondly, both model estimation and separation between inefficiency and noise call for specific assumptions to be made about the distribution of either term. The most used distribution for the inefficiency term in the empirical work is the half-normal distribution. This distribution makes the majority of the firms almost completely efficient, though there is no theoretical reason that prevents the inefficiency to be distributed otherwise.[17] Finally, it might be incorrect to assume that the inefficiency is independent from the regressors: if a firm knows its efficiency level, this could affect its input choices.[18]

The preceding problems, which appear under the cross-section stochastic methodology, are potentially solvable using panel data. The first drawback can be handled if T (the number of observations on each firm) is large enough. However, this final benefit of having access to panel data can be overstated since in practice many panels are relatively short (Kumbhakar and Lovell, 2000). Second, having access to panel data allows the researcher to avoid any assumption about the distribution of the inefficiency by, instead, assuming that firms' inefficiency is constant over time. Finally, not all panel data estimation techniques require the assumption of independence of the technical inefficiency term from the regressors. Basically this kind of models can be derived using two different deterministic estimation techniques: fixed-effects model and random-effects model. The fixed-effects model does not require the assumption of independence between the inefficiency term and the regressors, but at the cost of not allowing the inclusion of constant regressors (which are likely to appear in the utilities sector). In the presence of time invariant attributes of the firms that are omitted from the model, these would be captured in the fixed effects, mixing with the (in)efficiency term, when they should be classified otherwise. The random-effects model, in turn, allows the inclusion of time invariant regressors in the model, although at the cost of assuming that the inefficiency term is independent from the regressors.[19]

Both the fixed-effects and random-effects models are deterministic, in the sense that all the differences between the firms' effects are denoted inefficiency. However, if the researcher is willing to assume some distribution of the efficiency term, and to assume independence between the efficiency effects and the regressors, a stochastic Maximum Likelihood estimate is feasible. This approach is widely used in empirical analysis (Kumbhakar and Lovell, 2000).

The fixed-effects and random-effects models assume that the inefficiency is constant over time, but this assumption can be relaxed. If one finds the assumption that inefficiency is time invariant untenable (and it becomes increasingly so as the number of time series observations becomes larger), some structure of how the inefficiency evolves across time could be imposed. One possibility is the Cornwell, Schmidt and Sickles (1990) specification, which allow the individual effect to evolve over time as a quadratic function ($u_{i,t} = \gamma_{i,1} + \gamma_{i,2} t + \gamma_{i,3} t^2$). That is, the inefficiency term is a quadratic function of time, but the form is not the same across firms. The Cornwell et al. specification is very flexible, but is very demanding in terms of data. When the sample is not big enough other specifications have been proposed. For example, Battese and Coelli (1992) specify the inefficiency as an exponential function ($u_{it} = \exp[-\eta(t-T_i)]u_i$, where η is the only parameter to be estimated). In this specification, if η is positive then the model shows decreasing inefficiency effects, while if η is negative the inefficiency effects are increasing (Coelli et al. 1998). A disadvantage of this specification is that the ordering of the firms according to the technical inefficiency effects is the same at all time periods. The main advantage is that it is less data demanding than the Cornwell et al. (1990) model.

When panel data is available, frontier estimation methods (both parametric and non-parametric) can be used to obtain estimates of total factor productivity (TFP) growth. As Coelli et al. (1998) express, some of the advantages of following a frontier approach to TFP growth are that it does not require price data nor behavioral assumptions, it does not assume that all firms are fully efficient, and it allows the decomposition of TFP measures into technical change and technical efficiency change. This would permit the regulator the estimation of both components of an efficiency gain: gains from shifts in the frontier (technical change) and gains at the firm level reflecting a catching up effect (technical efficiency change). This decomposition is specially useful while setting the X factor in a regulatory framework of price cap and RPI-X regulation.

In all econometric panel data models technological change can be estimated by including a time trend (and eventually its square) in the regressor vector. The inclusion of a time trend reflects what is known as Hicks neutral technical change. That is, the intercept of the function shifts but the slope does not.[20]

II.6. The choice of variables

A frontier model has two parts: the “core” of the model, and the environmental variables. The (theoretically determined) core is formed by the inputs, in a production function approach, or the outputs and the input prices, in a cost function approach. The role of the environmental variables is to capture external factors that might influence the firms’ performance and are not directly controllable by them. Some examples of environmental variables include ownership differences, such as public/private, and location characteristics (see Fried, Schmidt and Yaisawarng, 1995). Most of the efficiency literature fails to recognize this decomposition of frontier models in regulated utilities.

As stated above, the initial specification for the core of the model is subject to theoretical considerations[21], and should be accepted or rejected as a whole, implying that it might be the case that some non-significant variables remain in the final model.

Environmental variables, on the other hand, are subject to different considerations: since these are not theoretically determined, they will be included in the final model only if they are statistically significant. The strategy would be to begin with an over-parameterized model followed by a stepwise regression procedure, to ensure that all the non-significant environmental variables are dropped from the final model. However, special care should be taken as regards the selection of environmental variables to be included in the initial (over-parameterized) model:

(i) In the case of ownership, for example, its inclusion as an explanatory variable gives information on the differences in efficiency for each ownership type. A set of dummy variables that measure these differences should not be included in a model intended for yardstick competition, for ownership effects would be netted out from the efficiency measures, thus punishing the firms belonging to the most efficient ownership type. If yardstick comparisons are to be made, the model should be estimated without these variables, and then the results (the relative efficiency measures) should be cross-checked with ownership information.

(ii) Geographical characteristics, on the other hand, are the kind of variables that should in general be included in the initial model, especially if the location of the firm is given by the concession contract (as is the usual case with regional monopolies). Because the firms cannot control their geographical environment, the efficiency measures should take into account that constraint.

(iii) Special attention must be taken in relation to the inclusion or not in the initial model of quality related variables. If quality standards do not exist, then the omission of quality variables in the model might cause some firms to appear with lower costs not because they are more efficient but because they provide a good or service of inferior quality. However, the regulator must have in mind that the inclusion of quality variables could result in quality standards above reasonable levels that would be passed on to the consumers through higher tariffs. If quality standards do exist, the optimal outcome results if the amount of potential fines is included in the computed costs.[22]

The idea behind the proposed stepwise procedure is that only environmental variables that are found to have a significant impact on costs should be recognized in the estimation of the efficient cost frontier. Non-significant variables do not help to explain variability in costs, and hence, should not be accepted by the regulator in an analysis that (precisely) attempts to establish the acceptable level of the costs incurred by the firms. This is the procedure followed, for example, by Stewart (1993) in his applied research done for the OFWAT.[23]

This, of course, does not mean that all significant variables have to be included in the model: only the significant *and* theoretically acceptable variables should be included. In practical terms this means that yardstick competition requires the regulator to recognize all those external factors that can affect costs (or the productive process). In other cases the possibility exists for the firm to engage in strategic behavior by explaining away firm specific inefficiencies as a state of nature (CRI, 1995): if, as we propose, an econometric approach were chosen to determine the final model, a firm could always find a variable that only it possessed, which, in statistical terms, would work as a dummy variable in the regressions, thus rendering it efficient –or more efficient.[24]

Of course, in many cases there are good reasons why some firms do not follow an efficient pattern, but once the regulators have done this initial sorting out, the burden of proof should be on the regulated companies. If they are indeed making the best effort to minimize cost, they should have enough information under their exclusive control to show that they are doing so and they should provide it to the regulator. This information should then be incorporated in any future work the regulators would use to compare companies, and become a component of standard informational requirements imposed on all companies (Crampes et al., 1997).

In this way, the initial model used as a yardstick is not so determinant, since the firms can challenge the proposed model until every part (firms and regulators) agree about the final model. In this sense, yardstick competition can be viewed as a “learning by doing” iterative process in which both firms and regulators learn while playing the game.

III. Consistency Conditions

A problem faced by regulators willing to apply frontier studies consists in the number of methods available for efficiency measurement of individual firms. The problem is far more serious if the different approaches give mutually inconsistent results.[25] The question then arises as to whether efficiency studies are empirically useful.

To overcome this problem, Bauer et al. (1998) propose a set of consistency conditions which must be met by the efficiency measures generated by the different methodologies, if these results are to be of some use to regulatory authorities.[26] For the comparison between approaches to make sense, the efficiency studies should refer to the same sample of firms (i.e. every methodology must consider the same firms and time period) and should make use of the same efficiency concept (see the section on regulatory choices above). The advantage provided by a consistency analysis is that the regulator can avoid the choice between approaches to efficiency measurement; plainly, the consistency conditions call for the use of several methodologies and for the cross-checking of results.

Specifically, the consistency conditions proposed by Bauer et al. (1998)[27] are:

(i) the efficiency measures generated by the different approaches should have similar means and standard deviations;

(ii) the different approaches should rank firms similarly;

(iii) the different approaches should identify, in general, the same firms as the "best" and the "worst";

(iv) the efficiency measures should be reasonably consistent with other performance measures;

(v) individual efficiency measures should be rather stable over time, i.e. should not vary significantly from one year to the other; and

(vi) the different measures should be reasonably consistent with the expected results from the industry, given the conditions under which it operates. In the particular case of regulated firms, for example, it is expected that those firms regulated under a price cap mechanism will be more efficient than those regulated under rate-of-return regulation.

Broadly speaking, the first three conditions determine the degree to which the different approaches are mutually consistent (i.e., if they are not met, individual efficiency measures generated by a single procedure would be somewhat subjective, and hence unreliable), whereas the remaining conditions establish the degree to which the different efficiency measures are consistent with reality. In other words, the first three conditions say if the different approaches will give the same answers to the regulators, while the last three conditions say if it is likely that these answers are correct.

If internal consistency is achieved (conditions (i) to (iii) are verified) the regulator can be confident that the figures (scores) obtained from the efficiency analysis are correct, and thus may proceed directly to setting an X factor for every firm under study.

If condition (i) is not met, but conditions (ii) and (iii) are, the regulator still has a rough ordering of the firms by their efficiency levels at hand, and therefore can discriminate the X factor by firms, starting from a common figure for this factor (perhaps one provided by a TFP growth study). Indeed, identifying the rough ordering of efficiency levels by firms is usually more important for regulatory policy decisions than measuring the level of efficiency itself (Bauer et al., 1998). OFWAT (1998), for instance, makes a discrimination of the X factor according to the following banding convention:

- Band A: well below predicted expenditure (less than 85% of C)
- Band B: below predicted expenditure (85-95% of C)
- Band C: around predicted expenditure (within 5% of C)
- Band D: above predicted expenditure (105-115% of C)
- Band E: well above predicted expenditure (more than 115% of C)

where C is the estimated cost obtained from an Ordinary Least Squares regression; so, the bands are constructed using an average function, not a frontier. This procedure has the advantage that it is not sensitive to the presence of outliers (if the bands were constructed as a distance from the best practice, a best performance that was in fact an outlier would distort the banding allocation of the other firms).

If nor the first nor the second consistency condition are met, but the third condition is verified (consistency in identifying best and worse performers), it would still be possible to use an alternative approach: to publish the results. This is the approach followed in the UK in the water and electricity sectors. The idea is to inform the users and allow them to compare prices and services across regions and give them a reason to put pressure on their own operator if it is not performing well.

The last three consistency conditions would be like “external criteria” for the evaluation of the different approaches. They can also be useful to choose between methodologies if there is no agreement among them. For example, it is often the case that parametric methods are consistent with each other, as are non-parametric methods, but there is lack of consistency between parametric and non-parametric approaches.[28] In this situation, conditions (iv) to (vi) could help in establishing which approach gives more correct answers, thus discarding every other methodology on the basis of a sound argument of inconsistency.

IV. Conclusions and Suggestions

In this paper we have dealt with the empirical application of the theoretical concepts developed by the efficiency measurement literature. We have considered a number of choices the regulator has to face when performing an efficiency analysis, and we have thoroughly discussed the regulatory implications of each particular choice.

We are now able to propose an efficiency measurement procedure that takes into account every applied consideration made in this work. This procedure involves the following steps:

(i) identify a set of comparable firms;

(ii) construct the theoretical core of the model: this step involves the selection of the kind of relationship that will be estimated (cost or production function), which has an implicit choice about the relevant efficiency concept; it also involves the definition of which variables are outputs and which are inputs;

(iii) select all the environmental variables that could potentially affect performance;

(iv) regress the initial model and follow a stepwise procedure to ensure that all the non-significant environmental variables are dropped from the final model;

(v) run a DEA model with the inputs, outputs and environmental variables selected in previous steps (final model), to identify efficient and inefficient firms;[29]

(vi) regress the final model, including a dummy variable which takes a value of one if the firm is found efficient in step (v), and zero otherwise;

(vii) apply the consistency condition analysis.

Once the regulator has completed this procedure, and is confident about her results, she can send the efficiency evaluation to each regulated firm, and invite responses from them. In this way, regulators can seek the involvement of the firms in the benchmarking process to ensure that the data on which the analysis is based is reliable and that the results are comprehensible and justifiable. Yardstick competition would then result in a “learning by doing” iterative process in which both firms and regulators learn while playing the game.

Notes

[1] See the discussion in Green and Rodríguez Pardina (1999).

[2] Technical efficiency can be further decomposed into "pure" technical efficiency, scale efficiency and congestion efficiency, as suggested in Färe et al. (1985), Pollitt (1995) and Coelli et al. (1998).

[3] The Spanish electricity distribution sector case is analysed by Grifell-Tatjé and Lovell (2000), who compare actual performance not against typical best practice standards, but against ideal engineering standards established by an international consultant. Unexpectedly (for us) they find that managers are more cost efficient than the ideal practice developed by the consultant.

[4] There are other types of functions that can be estimated (e.g. revenue function, profit function). However, cost and production functions are the most common and we only deal with these in this paper.

[5] In econometrics applications, if one wishes to conduct separate estimations on both types of inefficiency it is necessary to make some additional assumptions. In mathematical programming applications, it is necessary to run two separate programs for each firm: one to estimate technical efficiency and another for overall efficiency; allocative efficiency comes as a residual.

[6] In an econometric setting, an additional advantage stemming from the use of cost functions has to do with their flexibility to adapt to situations in which more than one output is produced. The analysis of multiproduct firms is straightforward in linear programming applications, even in the context of production relationships.

[7] Besides, in public firms, prices may be neither available nor reliable (Charnes, Cooper and Rhodes, 1978).

[8] This advantage only comes at a cost: neglecting prices, one can no longer estimate allocative efficiency.

[9] See Vanden Eeckaut et al. (1993), Ramos and Sousa (1998) and DTe (2000).

[10] This problem is more important with DEA models with variable returns to scale than in models assuming constant returns to scale.

[11] This problem is analogous to estimating a production function when output is exogenous.

[12] Other flexible functional forms are the generalized Leontief and generalized Cobb-Douglas. Guilkey, Lovell and Sickles (1983) compare all of them and conclude that the translog form performs at least as well as the other two and provides a dependable approximation to reality provided reality is not too complex.

[13] Though theoretically there have been advances in the development of non-parametric stochastic frontiers -the stochastic DEA models proposed by Land, Lovell and Thore (1993) and Olsen and Petersen (1995)-, in practice the mathematical programming is largely nonstochastic (Kumbhakar and Lovell, 2000).

[14] In a production function approach the inefficiency term is non-positive.

[15] OFWAT called the deterministic approach "regression analysis", but the main idea is the same.

[16] OFWAT (1998), however, recognised "that the differences between predicted and actual expenditures, even after adjustment for specific factors did not translate directly to differences in efficiency [...]. Therefore the approach adopted was to set company specific efficiency targets that would move individual company expenditure towards those of the best performers, over a five-year period. The amount of movement was taken to be around 25%-35% of the differences in predicted costs."

[17] A common criticism of the stochastic frontier method is that there is no a priori justification for the selection of any particular distribution form for the technical inefficiency effects (Coelli et al., 1998). Some authors attempted to address this criticism by specifying more general distributional forms, such as the truncated normal distribution (Stevenson, 1980) and the two-parameter gamma (Greene, 1990). However, the ultimate question is: do distributional assumptions matter? In an attempt to answer this question, Kumbhakar and Lovell (2000) find that, though sample mean efficiencies could be sensitive to the distribution assigned to the one-side error component, it is not clear whether the ranking of producers by their efficiency scores, or the composition of the top and bottom efficiency score deciles, is sensitive to distributional assumptions.

[18] If the regulator monitors the relative efficiency of the firms across time and adopts the procedure of submitting the results of the efficiency analysis to the firm for discussion, then it becomes more likely for this assumption to be violated.

[19] See footnote 18.

[20] That is, the marginal rate of substitution does not change. In a production function model the non-neutral technical change can be calculated including the interaction terms between inputs and time.

[21] The applied literature is a good starting point in the identification of the theoretical variables to be included in the core of the model. A survey of this literature is available from the authors on request.

[22] The water regulator in UK, for example, makes a strong case against financing discretionary quality improvements through higher prices, and adds that though in their response to the companies' market research some customers have said they would like to see improvements in levels of service, they have shown considerable resistance to pay higher prices for these improvements. Customers on lower income brackets encounter particular difficulties in paying higher prices and, therefore, the regulator will only make provision for enhanced service standards in future price limits where there is very clear evidence, across the whole spectrum of customers, of willingness to pay (OFWAT, 1994).

[23] A similar procedure can be found in Pollit (1995), who suggests that regression analysis be used to test the significance of the variables considered, in order to keep the number of variables as low as possible in DEA applications. Another applied work that recommends the use of regression techniques to identify cost drivers is DTe (2000), though a technique other than econometric tests is finally employed for model selection, due to the small size of the sample (which could produce misleading results in regression analysis). Kittelsen (1999) applies an stepwise procedure to discard some variables (inputs and outputs) from his model of the Norwegian Electricity Distribution Utilities.

[24] In a DEA setting a firm could make itself appear as more efficient by including additional environmental variables, because it would be difficult to find comparable firms in the set when an increasing number of dimensions is considered in the analysis (and not because it is actually efficient).

[25] Weyman-Jones (1992, p. 440) warns about the likelihood of regulatory debates being taken to the legal arena whenever the regulator and the firms disagree on the correct methodology used in efficiency measurement.

[26] Although there exists a vast literature on efficiency measurement in the utilities sector, few studies try to compare the efficiency measures obtained with the different approaches. Among them are the works of Pollitt (1995), Ray and Murkherjee (1995) and Burns and Weyman-Jones (1996). Neither of these authors, however, makes a consistency analysis as formal as the one in Bauer et al. (1998). Kittelsen (1999), in an applied paper on the regulation of the Norwegian electricity distribution utilities, states as a condition to apply DEA yardstick competition that the results be validated by statistical tests and *compared with other econometric methods*.

[27] The paper by Bauer et al. (1998) deals with the consistency of efficiency measures in the U.S. banking sector and finds mixed evidence as regards the fulfilment of the consistency conditions. Rodriguez Pardina et al. (1999) in a study of the electricity distribution sector in South America analyzed the set of conditions proposed by Bauer et al. (1998), finding that the different approaches are consistent in their means, rankings and identification of the same firms as the "best" and the "worst".

[28] See Bauer et al. (1998) and Rodriguez Pardina et al. (1999).

[29] As suggested in Arnold et al. (1996) and Bardhan et al. (1998).

References

Aigner, D., Lovell, C. y Schmidt, P. (1977), "Formulation and Estimation of Stochastic Frontier Production Function Models". Journal of Econometrics, Vol. 6, 21-37.

Ali, A. y Seiford, L. (1993), "The Mathematical Programming Approach to Efficiency Analysis". En Fried, H., Lovell, C.A.K. y Schmidt, S. *The Measurement of Productive Efficiency.* Oxford University Press.

Arnold, V., Bardhan, W., Cooper, W. y Kumbhakar, S. (1996), "New Uses of DEA and Statistical Regressions for Efficiency Evaluation and Estimation -With an Illustrative Application to Public Secondary Schools in Texas". Annals of Operations Research 66, 255-278.

Bardhan, W., Cooper, W. y Kumbhakar, S. (1998), "A Simulation Study of Joint Uses of Data Envelopment Analysis and Statistical Regressions for Production Function Estimation and Efficiency Evaluation". Journal of Productivity Analysis 9, 249-278.

Battese, G. y Coelli, T. (1992), "Frontier Production Functions, Technical Efficiency and Panel Data: With Application to Paddy Farmers in India". Journal of Productivity Analysis 3, 153-169.

Bauer, P., Berger, A., Ferrier, G. y Humphrey, D. (1998), "Consistency Conditions for Regulatory Analysis of Financial Institutions: A Comparison of Frontier Efficiency Methods". Journal of Economics and Business, 50, 85-114.

Bosworth, D., Stoneman, P. y Thanassoulis, E. (1996), "The Measurement of Comparative Total Efficiency in the Sewerage and Water Industry: An Exploratory Study". Report to and commissioned by the Office of Water Service, UK, Octubre.

Burns, P. y Weyman-Jones, T. (1996), "Cost Functions and Cost Efficiency in Electricity Distribution: A Stochastic Frontier Approach". Bulletin of Economic Research*,* 48,1.

Charnes, A., Cooper, W. y Rhodes, E. (1978), "Measuring the Efficiency of Decision Making Units". European Journal of Operational Research, 2 (6), 429-444.

Coelli, T., Prasada Rao, D. y Battese, G. (1998), "An Introduction to Efficiency and Productivity Analysis". Kluwer Academic Publishers.

Cornwell, C., Schmidt, P., y Sickles R. (1990), "Production Frontiers with Cross-Sectional and Time Series Variation in Efficiency Levels". Journal of Econometrics, Vol. 46, 185-200.

Crampes, C., Diette, N. y Estache, A. (1997), "What Could Regulators Learn from Yardstick Competition? Lessons for Brazil's Water and Sanitation Sector", Mimeo, The World Bank.

CRI (1995), "Yardstick Competition in UK Regulatory Processes". Centre for Regulated Industries, The World Bank, Junio.

DTe (2000), “Choice of Model and Availability of Data for the Efficiency Analysis of Dutch Network and Supply Businesses in the Electricity Sector”. Background Report, Netherlands Electricity Regulatory Service, Febrero 2000.

Färe, R., Grosskopf, S. y Lovell, C.A.K. (1985), “The Measurement of Efficiency of Production”. Boston, MA: Kluwer-Nijhoff Publishing.

Farrell, M. (1957), “The Measurement of Productive Efficiency”. Journal of the Royal Statistical Society, Series A, Part III, Vol. 120, 253:281.

Fried, H., Schmidt, S. y Yaisawarng, S. (1995), “Incorporating the Operating Environment into a Measure of Technical Efficiency”. Mimeo, Union College, Schenectady.

Green, R. y Rodriguez Pardina, M. (1999), “Resetting Price Controls for Privatized Utilities. A Manual for Regulators”. EDI Development Studies, Economic Development Institute, The World Bank, Washington D.C.

Greene, W. (1990), “A Gamma-Distributed Stochastic Frontier Model”. Journal of Econometrics, 46, 141-164.

Grifell-Tatjé, E. y Lovell, C.A.K. (2000), “The Managers versus the Consultants”. Mimeo.

Guilkey, D., Lovell, C. y Sickles, R. (1983), “A Comparison of the Performance of Three Flexible Functional Forms”. International Economics Review, 24 (3), Octubre, 591-616.

Kittelsen, S. (1999), “Using DEA to Regulate Norwegian Electricity Distribution Utilities”. Presentación en el 6th European Workshop on Efficiency and Productivity Analysis, Copenhagen.

Kumbhakar, S. y Lovell, C.A.K. (2000), “Stochastic Frontier Analysis”. Cambridge University Press.

Land, K., Lovell, C.A.K. y Thore, S. (1993), “Chance-Constrained Data Envelopment Analysis”. Managerial and Decisions Economics, 14, 541-554.

Lovell, C.A.K. (1993), “Production Frontiers and Productive Efficiency”. En Fried, H., Lovell, C. y Schmidt, S. *The Measurement of Productive Efficiency.* Oxford University Press.

Meeusen, W. y van de Broeck, J. (1977), “Efficiency estimation from Cobb-Douglas production functions with composed error”. International Economic Review, Vol. 18, Nº 2, Junio, 435-444.

OFWAT (1994), “Setting Price Limits for Water and Sewerage Services. The Framework and Approach to the 1994 Periodic Review”, Office of Water Services, Birmingham, UK.

OFWAT (1998), “Assessing the Scope for Future Improvements in Water Company Efficiency: A Technical Paper”. Office of Water Services, Birmingham, UK, Junio.

Olsen, O. y Petersen, N. (1995), “Chance Constrained Efficiency Evaluation”. Management Science, 41, 442-457.

Pestieu, P. y Tulkens, H. (1990), “Assessing the Performance of Public Sector Activities: Some Recent Evidence from the Productive Efficiency Viewpoint”. Discussion Paper Nº9060, CORE, Université Catholique de Louvain, Belgium.

Pollitt, M. (1995), “Ownership and Performance in Electric Utilities: the International Evidence on Privatization and Efficiency”. Oxford University Press.

Ramos, F. y Sousa, M. (1998), “Eficiência Técnica e Retornos de Escala na Produçao de Serviços Públicos Municipais: Uma Avaliaçao Nao-Paramétrica dos Custos Associados à Descentralizaçao Política no Brasil”, XX Encontro Brasileiro de Econometria, Vitória, Espírito Santo, Brasil, Diciembre 1998.

Ray, S. y Mukherjee, K. (1995), "Comparing Parametric and Non-Parametric Measures of Efficiency: a Reexamination of the Christensen-Green Data". Journal of Quantitative Economics, Vol 11, No. 1, Enero.

Reiter, H., McCarthy, S. y Harkaway, P. (1999), "Implications of Mergers and Acquisitions in Gas and Electric Markets: The Role of Yardstick Competition in Merger Analysis". Quarterly Bulletin, Vol. 20, Nº2, 193-199.

Rodríguez Pardina, M., Rossi, M. y Ruzzier, C. (1999), "Consistency Conditions: Efficiency Measures for the Electricity Distribution Sector in South America". CEER Working Paper Nº5, Mayo.

Schmidt, P. y Sickles, R. (1984), "Production Frontiers and Panel Data". Journal of Business & Economic Statistics, 2, Octubre, 367-374.

Shleifer, A. (1985), "A Theory of Yardstick Competition". Rand Journal of Economics, Vol. 16, 3, Autumn, 319-327.

Stevenson, R. (1980), "Likelihood Functions for Generalised Stochastic Frontier Estimation". Journal of Econometrics, Vol. 13, 57-66.

Stewart, M. (1993), "Modelling Water Costs 1992-93: Further Research into the Impact of Operating Conditions on Company Costs." OFWAT Research Paper Number 2, Diciembre.

Vanden Eeckaut, P., Tulkens, H., y Jamar, M. (1993), "Cost Efficiency in Belgian Municipalities". En Fried, H., Lovell, C. y Schmidt, S. *The Measurement of Productive Efficiency.* Oxford University Press.

Weyman-Jones, T. (1992), "Problems of Yardstick Regulation in Electricity Distribution". En Bishop, Kay and Mayer. *The regulatory challenge.* Oxford University Press.

Yunos, J. y Hawdon, D. (1997), "The Efficiency of the National Electricity Board in Malaysia: an Intercountry Comparison." Energy Economics, 19, 255-269.