Topics in efficiency benchmarking of energy networks: Choosing the model and explaining the results

Report prepared for

The Netherlands Authority for Consumers and Markets

15 December 2017

Denis Lawrence, John Fallon, Michael Cunningham, Valentin Zelenyuk, Joseph Hirschberg

Economic Insights Pty Ltd
Ph +61 2 6496 4005 Mobile +61 438 299 811
WEB www.economicinsights.com.au
ABN 52 060 723 631

CONTENTS

EXECUTIVE SUMMARY

Benchmarking refers to measuring the efficiency of businesses in comparison to a point of reference, usually the best observed practice. The main focus of this report is benchmarking via data envelopment analysis (DEA). This report addresses the topics of

- choosing a benchmarking method and then a preferred benchmarking model, and
- explaining and evaluating the results of DEA benchmarking analysis.

Although many approaches are discussed in the report, it needs to be recognised that feasibility and resource limitations will influence the most ideal or optimal approach that can be implemented in practice.

Benchmarking in regulation

Benchmarking methods are widely and increasingly used in regulation frameworks for energy utilities, and are either used as part of a specific ‘yardstick regulation’ framework for setting regulated revenues, or alternatively, may be used more broadly in combination with other methods for assessing the efficient cost of supply. The role of benchmarking in regulation is to provide incentives for businesses to improve their efficiency and ultimately reach best practice. While this report focuses on data envelopment analysis (DEA), alternative benchmarking methods include: (i) multilateral TFP indexes; (ii) corrected ordinary least squares (COLS); (iii) stochastic frontier analysis (SFA).

While efficiency benchmarking is particularly useful for reducing the asymmetry of information between the regulator and the regulated businesses, the reliability of the efficiency estimates for businesses depends importantly on the ability to control for firm-specific factors when making comparisons. Efficiency analysis from a short-term perspective may not fully capture some of the important aspects of efficiency in the use of long-lived assets.

DEA Methods

The conventional DEA model can be expressed in either the envelopment form or the multiplier form, and while these two approaches yield the same efficiency scores, they each provide somewhat different additional information. For this reason it is useful to use both approaches. If input prices are available, the DEA cost efficiency model can be used. This yields estimates of cost efficiency rather than technical efficiency. When the cost efficiency model is used in conjunction with the technical efficiency model, this enables a better understanding of the nature of inefficiencies.

Statistical bootstrapping can be used with DEA models to estimate confidence intervals for each efficiency score estimate for each firm in the sample. In small samples the confidence intervals may be quite wide, however, an understanding of the reliability of the efficiency estimates, and how they vary between the firms in the sample, can be useful information.

The assumptions made in a DEA study in relation to returns to scale have an important bearing on the efficiency estimates obtained. Several alternative assumptions are available.

The most standard assumptions are constant returns to scale (CRS) or variable returns to scale (VRS) or an intermediate form. However, while VRS is the most flexible of these alternatives, it is more restrictive than can be strictly justified on the grounds of economic theory. An alternative is the Petersen (1990) approach in which returns to scale are not constrained to be convex.

The Farrell efficiency score obtained from the conventional DEA program is based on radial contraction of inputs or radial expansion of outputs (i.e. preserving the mix). Because of this, there may be slacks in the use of one or more inputs, or in under-producing one or more outputs. If slacks are present, a Farrell-efficient mix of inputs and outputs will not be economically efficient. The identification of slacks and calculation of efficiency scores that take them into account is important, but it appears not often done in energy utility benchmarking.

In utility regulation contexts, the input-oriented efficiency measures are most commonly used, because output is not usually a discretionary variable for these businesses. However, in some circumstances it may be outputs and inputs are both in part controllable. For example, output quality may be in the control of a utility. Alternative orientation assumptions include: additive models, hyperbolic measures of efficiency, non-oriented Russell measures, the geometric distance function and directional distance functions.

The weights in the DEA multiplier program, which are interpreted as shadow prices, differ for each firm. The amount of variation in these weights can be problematic in some circumstances. One approach to dealing with this is to impose subjective weight restrictions. Another approach is to require the model to have a single common set of weights for all firms. Often in regulatory applications of DEA a single input is used (total cost). This can be viewed as another way of imposing restrictions on input weights.

Several other extensions to the DEA model are briefly surveyed including: (i) latent class models; (ii) dynamic DEA; (iii) free disposal hull; and (iv) stochastic nonparametric frontiers. Most of these methods have been used in energy network benchmarking, and the last of these methods has been used for regulatory benchmarking of electricity distribution in Finland. Also reviewed are methods for dealing with operating environment variables. Several approaches are discussed, but the two-stage approach has a number of advantages and is the most popular method. It involves regressing estimated DEA efficiency scores against operating environment factors in a second-round regression analysis. Special methods or censored regression have been developed to ensure that this can be done robustly.

Regulation of TSOs and Benchmarking

This report reviews the methods of regulating electricity and gas TSOs used in a number of countries and the benchmarking methods and applications used. The countries surveyed include a number of European countries (Finland, Germany, Ireland, the Netherlands, Portugal, Sweden and the United Kingdom) and several countries in other parts of the World (Japan, Brazil, New Zealand, the USA and Australia). Most of these regulators, and almost all of the European regulators, use some form of benchmarking as part of the analysis they undertake. Various benchmarking methods are used and for the most part unique to each regulator. In some jurisdictions benchmarking is not used as much for electricity and gas

TSOs compared to the regulation of distribution networks, in part because state-owned monopolies are more common in energy transmission, and in part because there may be fewer comparators. The USA does not appear to make substantial use of benchmarking for electricity and gas TSOs. Australia is currently developing its benchmarking framework for electricity TSOs.

Selecting a Preferred Model

The model selection process in DEA has parallels to model selection in regression, particularly so given the recent developments in statistical foundations for DEA (including bootstrapping). Alternatively, there are a number of useful regression-based methods that can be used to assist model selection in DEA.

At the outset of a benchmarking exercise it is necessary to decide which assumption about returns to scale is most plausible for the industry at hand. This choice is important in DEA, because it influences the estimated measures of efficiency, and if the wrong assumption is made the estimates will generally be inconsistent. In the DEA context, Simar and Wilson (2002) have developed tests that can be employed to decide on the preferred returns to scale assumption, based on scale efficiency measures and using bootstrap methods. Applying statistical tests within a nonparametric framework using bootstrapping usually requires large samples, which may restrict the usefulness of this method in energy TSO benchmarking, unless sample sizes can be increased.

In DEA, parsimony in the inputs and outputs used is particularly important. This involves eliminating unimportant variables and aggregating variables where it is feasible to do so. As more variables are included in a DEA model the ability to discriminate between truly efficient and inefficient firms is reduced because more firms appear efficient purely because of the increase in dimensionality. Two of the approaches used for variable reduction that are examined include:

- Stepwise DEA, which is an adaptation of the stepwise regression procedure to DEA. This is an inexact procedure and requires large samples.
- Bootstrap tests, which are statistical tests that rely on bootstrapping to estimate the distributions of the test statistics. Although methodologically sounder than stepwise DEA, it still relies on very large data samples.

In these circumstances, some or all of the following methodologies may be feasible methods for narrowing down the candidate variables and choosing the final inputs and outputs.

(1) Use of Industry expertise

(2) Other DEA-based variable selection methods including, *inter alia*:

   - The ‘efficiency contribution measure’ method of Pastor et al (2002), which involves comparing differently specified DEA models to determine the incremental effect of each variable on the efficiency measures of firms.
   - The regression-based approach of Ruggiero (2005), an iterative process beginning with a minimally specified DEA model, regressing the resulting efficiency score estimates on remaining candidate variables, and identifying any significant

variables that might be added to the model.

(3) Use of stochastic frontier analysis (SFA) or other econometric model as a preliminary analysis to identify the most relevant inputs and outputs.

(4) Widening the sample for the purpose of deriving the model specification, for example by including North and South American TSOs in that stage of the analysis

(5) Use of principal components analysis (PCA) to transform the set of original variables into a smaller group of derived variables that contain much of the information in the original variables, thereby reducing dimensionality with minimal loss of information.

DEA results can be particularly sensitive to outliers, and it is important to identify outlier observations. Most of the methods used to identify outliers are based on 'super-efficiency' measurement. This is the efficiency measure obtained for a firm when it is itself excluded from the set of comparator firms that define the efficiency frontier. These measures are not bounded by one, and efficient firms will usually have super-efficiencies that are greater than one, but vary from firm to firm. Outlier detection is a largely *ad hoc* procedure of excluding the firms with the highest super-efficiencies. Once outliers are determined there is a question about how to deal with them. Although some authorities recommend automatically eliminating them from the sample, it is advisable to firstly better understand what they represent.

The objectives of the regulatory framework are another consideration in selecting a preferred model. This is because the benchmarking model, and the targets generated by it, may have an influence on the incentives of regulated businesses. For example, the omission of certain variables may take away incentives that the regulator would like to maintain.

Testing the Model's Representativeness

Chapter 6 explores a number of approaches relevant to testing the representativeness and reliability of a DEA model. A general question in evaluating the results of a model is whether the efficiency scores and rankings obtained from the analysis are consistent with other available information, which can include previous benchmarking studies or the views of experts with more detailed knowledge of the operating practices of the businesses being compared. When results are inconsistent with other sources of information, then further analysis is warranted to understand the results in more detail so that the benchmarking model can be critically evaluated.

It will be particularly useful to carry out a similarly specified stochastic frontier analysis (SFA) to compare with the DEA results. If the two methods give quite different results for a particular TSO, then this may indicate that the DEA score for that business may be comparatively unreliable. Comparison of efficiency rankings obtained using DEA and SFA may also be instructive.

Other assessments that may be needed to ensure a proper basis for comparing the efficiencies of firms in the sample include:

- It is important to identify and take account of 'slacks', which are sources of inefficiency not taken into account in the conventional radial efficiency estimates.

Efficiency scores or rankings can be adjusted if there are slacks. This is important because slacks are usually quite common. Some firms may be incorrectly assessed as efficient if an adjustment is not made for slacks.

- Another test of reliability is to quantify and remove the estimated effect of 'sampling bias', which can lead to over-estimation of efficiency scores. However, we found that quite large samples are needed to obtain reliable estimates of the sampling bias, and therefore this kind of adjustment is unlikely to be feasible in the contexts of TSO benchmarking.
- Sensitivity analysis can be undertaken to ascertain how much data error would be needed to substantially change an efficiency score, or alter conclusions about whether a TSO is efficient or inefficient. This sheds light on the robustness of the DEA efficiency estimates, which can be important to their proper interpretation.
- The effects of operating environment factors on efficiency scores should also be quantified. A 'second-stage analysis' can be used for this purpose and efficiency scores can be adjusted for the effects of the exogenous operating environment factors.[1] It is important to adjust for the effects of these factors because they cannot be influenced by actions management can take and are not related to the performance of the firms. Second-stage analysis is the most accepted method of controlling for these influences.

Further Analysis

Methods of further analysis that can be undertaken to improve understanding of the DEA benchmarking results are discussed in chapter 7. Firstly, when DEA input-oriented technical efficiency analysis is undertaken together with cost efficiency analysis, the cost efficiency score can be decomposed into allocative and technical efficiency. This decomposition helps to explain the sources of inefficiency and is important information for TSO management because strategies for reducing technical inefficiency may differ from those needed to reduce allocative inefficiency. The second type of analysis discussed in chapter 7 is the calculation of the Malmquist productivity index to obtain estimates of total factor productivity changes for each TSO. This is important information for several reasons: productivity trends can be a useful diagnostic check on the benchmarking model; the performance of TSOs can be compared with their own past performance, and their efficiency gains be compared to those of other TSOs. There are also several useful ways in which changes in the Malmquist productivity index can be decomposed into separate explanatory factors, including technical change (or 'frontier shift'), changes in technical efficiency (or 'catch-up') and the effects of changes in output on scale efficiency.

[1] By 'exogenous' we mean that the operating environment factors are exogenous for the firm. Management can still make choices in how to deal with operating environment factors (which may be more or less effective), but these responses generally require resources to implement, so that differences in operating environments can affect the observed comparative productivity and cost efficiency of firms even when action is taken to mitigate their effects. The effect of operating environment factors is an empirical question.

A third useful type of analysis is the calculation of elasticities of substitution between inputs, or similar analysis, which quantifies, in economic terms, the technical characteristics of the estimated production possibilities set. This information can be compared to expert opinions on the characteristics of the technology, and to previous findings in the literature on the cost structure and marginal rates of transformation and substitution in energy networks.

Combining Models

It is usually desirable to use more than one benchmarking technique for the purpose of methodological cross-checking. If one model is not clearly superior to another then one approach is to combine the models in some way to obtain estimated efficiencies. Chapter 8 examines some of the methods for combining models. The Bayesian Model Averaging method is of particular interest. It is designed to take account of model uncertainty, which is often ignored, particularly when one model is chosen as a preferred model when there is an alternative model with a significant likelihood of being the better model. BMA is a method of model averaging that uses weights for each model based on the likelihood of that model being the 'true' model. The averaging may be of the estimated efficiency scores of the different models and/or the estimated probability distributions for the estimated efficiency scores. This method can be used to combine different DEA models, or to combine a DEA model with an econometric model such as SFA. The DEA and SFA approaches to efficiency measurement each have their own strengths and weaknesses. An approach that combines a preferred DEA model with a preferred SFA model may have merit and is well worth considering.

Benchmarking

Use of DEA efficiency scores for benchmarking purposes is discussed in chapter 9. One of these uses is setting targets for inputs given the anticipated levels of outputs. In yardstick regulation frameworks, price or revenue caps are usually based on the estimated efficient cost of supply, allowing for the time that may be needed to achieve efficiency. Implicit within DEA efficiency scores are targets for inputs, which are related to the efficient cost of supply via forecasts of demand and input prices. Information on the implied input targets is likely to be useful to the regulator when setting the regulatory controls, and may also be useful to businesses to translate the revenue or price caps into targets that are directly within their control.

A second use of DEA results for benchmarking purposes is identifying the efficient peers of inefficient TSOs. To become efficient the TSO may need to become more like its efficient peers. Therefore, once the efficient peers have been identified, a more detailed comparison can be undertaken, as case studies, between the inefficient TSO and its efficient peers to seek a better understanding of why those businesses are more efficient. The efficient peers can be seen as role models because they have a similar mix of inputs and outputs, and therefore similar operations, and what they do differently or better than the inefficient business may shed light on the reasons for its inefficiency.

Chapter 9 also explains some graphical methods of comparing TSOs with each other, or with the projected mix of inputs and outputs implied by their efficiency score. The radar diagram is explained as a particularly useful graphical tool for this purpose. Ranking of units can be useful for both descriptive and analytical purposes. The use of rankings and of ‘context dependent’ DEA are discussed as methods for identifying subgroups of TSOs that may be considered to be at similar efficiency levels, which is another perspective relevant to competition by comparison. Productivity growth rates of TSOs within like groups, or between sub-groups of TSOs can be compared. Formal tests of general hypotheses can be carried out, such as whether the extent of catch-up is greater among the least efficient firms, than among firms that are closer to the efficiency frontier. Monitoring productivity growth can also shed light on the effectiveness of the regulatory framework, including whether it is resulting in the efficiency gains that were expected at the time of the last revenue cap determination, and whether there is any correlation between the types of regulation framework and the productivity gains observed.

Further Topics

Chapter 10 describes a number of good practices in documenting benchmarking studies, largely drawn from guidelines issued by competition agencies on standards relating to the submission of economic evidence. These guidelines suggest that expert benchmarking reports should meet two overall aims. Firstly, they should be sufficiently thorough not only in relation to the documenting of data and methodologies in the final analysis, but also with regard to the process of reaching the final analysis, including both the reasoning processes and the quantitative investigation steps. Secondly, the presentation of the study should aim to give the reader an understanding of the key aspects of the analysis and results. For example, by identifying important features of the technology which explain the choices of variables used in the study; aspects of the dataset that have had an important bearing on the results; interpretations of quantitative results in terms of economic theory, and generally to explain and illustrate the results succinctly but effectively.

Chapter 10 also discusses the potential for using economic benchmarking frameworks in conjunction with individual firms’ more specific performance frameworks such as *key performance indicators* (KPIs) and *balanced scorecards* (BSCs). One issue is whether datasets gathered as part of the benchmarking exercise might assist firms to operationalize strategies to improve their overall economic efficiency in accordance with the objectives of the regulatory benchmarking framework. More generally, within a regulatory setting, to be fully effective the KPI or BSC frameworks need to be developed with an understanding of how performance in particular dimensions influences overall economic performance. These observations suggest that, ideally, business KPI frameworks designed to improve efficiency of particular activities or dimensions of business performance should be complementary to the effectiveness of the regulatory benchmarking framework. This aspect of how the benchmarking methods or data might be translated by firms into lower level strategic performance management may be worthy of consideration at the time data requirements are developed and benchmarking is carried out.

1 INTRODUCTION

Benchmarking refers to measuring the efficiency of businesses in comparison to a point of reference, usually the best observed practice. This report addresses the topic of choosing a benchmarking method and then a preferred benchmarking model. The main focus of this report is benchmarking via data envelopment analysis (DEA), although other benchmarking methods are also discussed briefly in section 2, and further in section 7 in the context of combining the results of different benchmarking models. This paper discusses a wide range of DEA techniques and methodologies with particular emphasis on those that can or have been used to analyze energy network efficiency. Techniques that have been used in benchmarking for energy network regulation are noted.

This report also discusses issues related to explaining and evaluating the results of DEA benchmarking analysis. Although logically these matters come after a benchmarking analysis is carried out, in practice the estimation of a benchmarking model is an iterative process involving successive rounds of evaluation and selection, and the steps involved in explaining the results of a model should be regarded as a part of the modelling process.

The report is structured as follows:

- Chapter 2 addresses the principles of regulation that are potentially important considerations in the choice of a preferred DEA model.
- Types of DEA methods that could be used or have been used in energy network benchmarking and described in chapter 3.
- Approaches to the regulation of electricity and gas TSOs in different jurisdictions, including applications of benchmarking methods, are addressed in chapter 4.
- Chapter 5 presents methodologies for choosing a preferred DEA model including, the assumptions to be made about scale economies, methods for selecting the final set of cost drivers and methods for detecting and managing outliers.
- Some available approaches for assessing the degree of reliability of the results of DEA analysis are discussed in chapter 6, as well as methods for adjusting efficiency estimates to make them more representative. The topics addressed include comparing results against other sources of information or other models, sensitivity analysis of the robustness of efficiency estimates, taking slacks into account, adjustments to estimated efficiency scores for bias, to take account of slacks, and to adjust scores for the effects of differences in operating environment characteristics.
- Chapter 7 discusses further analysis of benchmarking results to derive quantitative information beyond efficiency estimates, such as: the decomposition of cost efficiency into technical and allocative efficiency; calculation of the Malmquist productivity index to determine the change in productivity between periods and its decomposition into sources of productivity change; analysis of multipliers including calculation of elasticities of substitution between inputs;
- Chapter 8 discusses methods that can or have been used to combine the results of more than one acceptable model to derive reliable and representative efficiency

scores.

- The use of efficiency estimates in benchmarking is addressed in Chapter 9, including target setting for the inputs of benchmarked firms; detailed investigation of DEA-identified efficiency peers; graphical presentations of comparisons; ranking the firms in terms of efficiency and identification of sub-groups of like firms in terms of efficiency; and comparisons over time.
- Good practices in documenting the results of benchmarking studies; and the possible use of benchmarking studies by regulated businesses to inform the development of their own performance management systems are the subject of chapter 10.

2 THEORY OF USING DEA IN REGULATION

This section briefly reviews the theory behind the use of benchmarking in regulation, and particularly the use of nonparametric methods such as DEA. It also notes theoretical principles that are potentially important considerations in the choice of a preferred DEA model.

2.1 Benchmarking in Regulation

Benchmarking methods are commonly and increasingly used in regulation frameworks for energy utilities, and are generally used in one of two ways. The results of benchmarking may be used as part of a specific framework for setting performance targets for individual firms embodied in their revenue or price caps. Alternatively, the results may be used more broadly to provide information to the regulator and other stakeholders, including as a cross check against other methods of assessing the efficient cost of supply.

The first of these ways of using benchmarking is based on the notion of 'yardstick competition'. This is a method of regulation in which the allowed prices or revenues of one firm depend on the costs of similar firms. It thereby separates a firm's allowed prices from its own cost outcomes to provide strong efficiency incentives. This helps to address a key problem in economic regulation—firms have superior knowledge to the regulator of the technological possibilities and the efficient costs of supply. This uncertainty can prevent the regulator from achieving ideal (or 'first best') outcomes for consumers while at the same time ensuring the regulated firms have a reasonable likelihood of being financially viable.

In its earliest formulation (Shleifer, 1985), it was assumed that there were comparator firms of the same size producing exactly the same product, which could be used as yardsticks for each other. Setting the firm's price based on the cost outcome of the comparator ensured that any inefficient cost choice by a firm would not influence the price it received, and since that price is entirely exogenous to the firm, it would have strong incentives to minimize cost in order to maximise profit. This notion was extended by Bogetoft (1997, p. 278) to the case of multiproduct firms that are heterogeneous in terms of scale and product mix. Bogetoft developed an agency-type model to show that in these circumstances ideal outcomes cannot be achieved and the compensation framework for the regulated firm needs to be devised to attain the best possible trade-off between incentivising cost efficiency while minimising the 'information rents' captured by those firms. Benchmarking and relative performance evaluation plays a key part in the best achievable regulatory scheme. In Bogetoft's analysis, the optimal compensation scheme will involve some compromise between the efficient external benchmarks and firm's own cost outcomes. This means that the regulator needs to choose a weight to assign to the best practice norm (the balance of the weight being assigned to the firm's own cost outcomes).

Bogetoft also examined the merits of using DEA as a benchmarking tool, noting that it had gained popularity as a method for making performance comparisons. The study concluded that the implicit assumptions about technology within the DEA method, and its ability to take account of differences in the scale and product mixes of businesses, met the basic requirements of the yardstick regulation theory. However, other benchmarking methods (not considered) may also satisfy these requirements. Bogetoft recommended the use of DEA:

> DEA seems particularly well-suited for regulatory practice. First of all, it requires very little technological information a priori. Secondly it allows flexible non-parametric modeling of multiple-input multiple-output production processes in contrast to the stylized processes typically considered in the incentive regulation literature. Thirdly, DEA-based cost estimates are conservative or cautious, because they are based on an inner (minimal extrapolation) approximation of the production possibilities. (Bogetoft, 1997, p. 278)

While the usefulness of benchmarking in reducing the asymmetry of information between the regulator and the regulated businesses is clear, its effectiveness, and the weight that can be given to the results, relies heavily on the ability to control for firm-specific factors when making comparisons. Comparative costs or cost variations need to be "normalized for exogenous differences in firm attributes to develop normalized benchmarks costs … [which] can then be used by the regulator in a yardstick framework or in other ways to reduce its information disadvantage, allowing it to use high powered incentive mechanisms without incurring the cost of excessive rents" (Joskow, 2006, p. 14). Differences between utilities that are outside management control may include for example, topography, climate, customer density or regional input cost differences.[2] These exogenous factors produce heterogeneity in the underlying technological possibilities of the firms.

In some cases exogenous factors of this kind may be unknown or not measured, and thus cannot be controlled for. This is the general problem of unobserved firm-specific heterogeneity. By implication, the assumption that all firm-specific effects are entirely due to differences in technical inefficiency may be incorrect. Some methods have been developed relatively recently to better deal with this problem, as discussed in section 3.7.1.

Another point to note is that an assessment of efficiency in the use of capital inputs obtained in a benchmarking study carried out at a point in time may not fully capture some of the important aspects of making such an assessment for long-lived assets. Paulun et al (2008) note that an apparent sub-optimality of existing physical assets from a short-term perspective need not reflect inefficiency. It may arise because past network planning decisions were made in the absence of certainty about future market developments, or it may be that the optimality of the infrastructure can only be fully assessed from a long-term perspective.

The role of benchmarking in regulation is to provide incentives for businesses to improve their efficiency and ultimately reach best practice. In the short-run, efficiency targets need not be referenced against the best practice utility. In some cases they may be referenced against the average utility or an intermediate standard such as the margin of the top quartile (Lowry and Getachew, 2009, p. 1323). In the UK, the electricity regulator has used a target of the 75th percentile while the water regulator has also placed emphasis on the efficiency frontier

[2] By 'outside management control' we mean that the operating environment factors are exogenous for the firm. Management can still make choices in how to deal with operating environment factors (which may be more or less effective), but these responses generally require resources to implement, so that differences in operating environments can affect the observed comparative productivity and cost efficiency of firms even when action is taken to mitigate their effects. The effect of operating environment factors is an empirical question.

(Dassler et al., 2006). There is some debate about which standard should be employed and whether the concept of a 'normal' rate-of-return' only applies to firms with average efficiency[3] (see Kaufmann and Beardow, 2001; Lowry and Getachew, 2009; Shuttleworth, 2005; Tardiff, 2010). This debate highlights that care is needed to ensure that the use of frontier efficiency standards in regulatory compensation schemes (such as revenue caps) do not lead to unrealistically high or unachievable targets being set.

2.2 Alternative Benchmarking Methods

DEA is one of several benchmarking methods used in economic regulation. The most important among the alternative methods are:

- *Multilateral TFP indexes*: an index number method of TFP calculation which permits invariant productivity comparisons between firms and over time via the overall sample average (Caves et al., 1982a);
- *Corrected ordinary least squares* (COLS): econometric analysis of production relationships in which the residuals are interpreted as measures of inefficiency, and the frontier is calculated by adding the largest positive or negative residual to the predicted values (depending on whether a cost function or a production function is being estimated);
- *Stochastic frontier analysis* (SFA): an econometric method for directly estimating efficiency frontiers that are subject to random disturbances.

Several studies have shown that there is often a lack of consistency in the results obtained using different benchmarking methods, particularly with relatively small data samples (Farsi et al., 2007, pp. 12–13). This should encourage rather than discourage the use of more than one method, because it may reduce the uncertainties, and "significant uncertainties in efficiency estimates could have important undesired consequences especially because in many cases the efficiency scores are directly used to reward/punish companies through regulation schemes such as price cap formulas" (Farsi et al., 2007, p. 13). Most of this report is focussed on non-parametric benchmarking methods such as DEA, but chapter 8 addresses the desirability of combining the results with those of other benchmarking methods such as stochastic frontier analysis. Alternatively, the use of several methods may provide corroboration of the results of a preferred model.

2.3 Good benchmarking practices

Haney and Pollitt (2012) suggest the following principles should be followed in efficiency benchmarking analysis, which they attribute to Knox Lovell:

[3] It is argued that the 'normal' rate-of-return on assets, or weighted average cost of capital—which is usually embodied in the cost benchmarks—is based on the average returns of firms with comparable risk. However, average returns, it is argued, are obtained by firms of average efficiency, whereas the most efficient firms obtain above-normal rates of return. This argument suggests that the efficient cost benchmark should embody an above average rate-of-return for firms with comparable risk.

- Use of frontier methods with enough variables to reflect the main feasible trade-offs
- A large high quality panel dataset
- Consistency with engineering knowledge about the underlying technology and ‘well behaved’ functional relationships
- Use of bootstrapping for confidence intervals
- Results for relative efficiencies should be consistent with industry experts’ views
- Appropriate operating environmental variables should be included in the analysis, and
- The efficiency analysis should demonstrate how it adds value.

3 DEA METHODS

This chapter discusses DEA and other nonparametric benchmarking methods that may be applicable and relevant to energy transmission benchmarking. It surveys a number of techniques and alternative formulations that that can be used with the DEA approach. The aim of this discussion is to identify a range of modelling techniques from which the preferred modelling method may be chosen. There is no intention to suggest that more sophisticated or recently developed techniques are to be preferred to simpler or more established methods. However, the review highlights some useful methodologies, with particular attention to those previously used in energy network benchmarking applications.

This chapter is structured as follows:

- Alternative programming methods for conventional DEA analysis and their different attributes are described in section 3.1.
- The use of bootstrapping to obtain confidence intervals for efficiency estimates is outlined in section 3.2.
- Section 3.3 discusses the different assumptions that can be made regarding returns-to-scale and the importance of those assumptions to the findings.
- Limitations of the Farrell efficiency measures obtained from conventional DEA models are discussed in section 3.4, as well as methods of identifying the subset of Farrell efficient firms that are fully economically efficient.
- Section 3.5 discusses the choices relating to output or input orientation, and various alternative or more general approaches. These methods all measure the efficiency of firms somewhat differently.
- Section 3.6 addresses the issue of controlling or limiting the multipliers or weights of the DEA model. This includes constraining the values the weight can take, requiring the firms to have a common set of weights, and implicit constraints arising from aggregating inputs into a measure of total cost
- Several variations on, or further developments of, the standard DEA model are briefly discussed in section 3.7, including latent class models, dynamic DEA, free disposal hull (FDH) and stochastic nonparametric frontiers
- Section 3.8 returns to the topic of taking operating environment variables into account via second-stage regression (introduced in our report ‘Selecting cost drivers’).

3.1 Alternative mathematical programming approaches

The basic DEA mathematical programming model of technical efficiency (whether input or output-oriented) involves solving a linear programming (LP) problem for each firm in the sample. It has two general formulations: the *multiplier* form and the *envelopment form.* This brief description focuses on the input-orientation case.

In the *multiplier form*, the problem is to find the values of a set of output and input weights for a firm *k* that essentially maximise its productivity (defined as the weighted sum of outputs divided by the weighted sum of inputs), although the weighted sum of inputs is normalised to equal 1. The weights must satisfy the constraint that, when applied to all other firms in the sample, the resulting productivity ratios are not greater than one. Thus the technical efficiency score of firm k relative to all other firms in the data sample is based on a set of weights chosen for firm *k*, which yield the highest feasible technical efficiency for firm *k*, and a different set of weights is found for each firm. The weights can be interpreted as normalised shadow prices (Coelli et al., 2005, p. 163).[4]

The *envelopment form* is the dual to the multiplier form, yielding the same efficiency scores. This mathematical program involves finding a set of non-negative peer weights (λ's) that minimise the Farrell efficiency score for firm k, (θ_k), subject to technology characterising constraints. In this way, the observed or DEA-estimated best-practice frontier is the smallest piecewise convex linear envelope that fits the data on inputs and outputs. The Farrell efficiency score represents the maximum proportion by which all inputs can be equiproportionately contracted such that the same set of outputs can still be produced (with the same technology). If $\theta_k = 1$, firm *k* is Farrell-efficient because the same output could not be produced with any small radial contraction of inputs. If $\theta_k < 1$, then firm k is considered as Farrell-inefficient because the same outputs could be produced with less inputs. The value of θ_k, when multiplied against firm k's actual inputs, projects the inputs onto the observed best-practice frontier. The projection point is defined as a convex combination of the inputs of the peer firms and the λ's represent the peer-weights.

The two forms are equivalent representations of the same production problem. The envelopment form is more commonly presented in economics applications while the multiplier form is more popular in the operations research and management science literature. The envelopment form provides information on peer DMUs, and the multiplier form provides other useful information, such as the shadow prices for the inputs and outputs. The multipliers can be used to calculate technical elasticities of factor substitution between inputs or between outputs, and marginal rates of transformation between inputs and outputs, which are related to ratios of the multipliers. For details of the calculation of substitution and transformation elasticities see Olesen and Petersen (2003) and Schmidtz and Tauchmann (2012). This topic is discussed further in section 7.4.

Given that the primal multiplier model and the dual envelopment model provide some different information, there is some benefit to computing both. The envelopment model multipliers provide direct information on the peer units and their relative weights. The weights from the multiplier model, or their ratios, can be scrutinised by experts to determine whether they are within reasonable ranges of values. If weight restrictions are imposed, then the multiplier form is usually more convenient.

The two models described above are for calculating technical efficiency. In addition, there is a cost-minimisation model, which involves solving for the cost minimising mix of inputs, given the set of input prices and the technology. Cost efficiency is defined as the minimum cost divided by the actual cost. This is usually estimated in conjunction with the input-

[4] This description is for the constant returns to scale (CRS) case. Additional restrictions on dual weights are made in the variable returns to scale (VRS) case.

oriented technical efficiency model, which together provide measures of technical efficiency, allocative efficiency and cost efficiency for each firm, when there is more than one input. An important benefit of cost-efficiency analysis (again, when there is more than one input) is that it enables targets to be developed for the changes in individual inputs needed for inefficient units to minimise cost. For example, an allocatively inefficient firm may require decreases of different proportions in inputs, or increasing one input and decreasing another, to achieve cost efficiency. This is potentially valuable information for the firms being benchmarked.

3.2 Bootstrapping DEA Results

Statistical bootstrapping is one approach to account for randomness of data and hence enable the use of statistical inference within the non–parametric DEA method. It has various applications in DEA, some of which are discussed in chapter 5 on 'selecting a preferred model'. This section briefly discusses methods of assessing the sensitivity of efficiency measures to variation in sampling, which can be used to shed light on the reliability of the efficiency estimates and correct for bias.

The bootstrapping method is based on the idea that the data sample is a random drawing from a larger population. Hence a DEA score obtained from that data sample is an estimate of the 'true' unknown efficiency, with some statistical uncertainty. In the absence of being able to draw more samples from the population, it is possible to randomly re-sample from the existing dataset (i.e. perform bootstrapping) to obtain information on the probability distributions of the DEA efficiency estimates. Bootstrapping is a well-established technique based on using a large number of samples, each consisting of data randomly drawn from the original dataset, applying DEA to each bootstrap sample, and calculating statistics such as means and standard deviations of the efficiency scores from the results.[5]

Under fairly broad assumptions about the underlying data generating process (including that the data sample is randomly drawn from a larger population), the DEA efficiency estimators are biased towards one, showing less inefficiency than when measured against the true (but unobserved) frontier. This is because they are defined with respect to the observed best practice frontier, based on the most efficient units in the data sample, which may not be fully efficient relative to the unobserved true frontier. The bootstrapping technique can be used to estimate the bias of efficiency estimates, which is a particular concern in small samples, and also estimate confidence intervals for efficiency comparisons, and other statistics that shed light on the reliability of efficiency scores.

Fried et al (2008, p. 59) observe that in relatively small samples the confidence intervals obtained are often "sufficiently wide to question the reliability of inferences drawn from such comparisons" between DMUs. They have been used in some of academic studies. See Hawdon (2003) for an application to the international gas industry and Jamasb et al (2008) for application to gas transmission companies. Hawdon found that whereas two DMUs may be estimated to have similar efficiency scores in conventional DEA, when bootstrapping is used, the efficiency score of one DMU may be found to be quite robust, whereas for the other

[5] Here the application is DEA, but bootstrapping can be used with other estimation techniques.

it may be quite unreliable. For this reason, he suggested that using conventional DEA scores in regulation without having regard to properly estimated confidence intervals can be problematic.

Confidence intervals on DEA scores are not widely used by regulators (Haney and Pollitt, 2012, p. 24). The reporting of wide confidence intervals in small-sample studies may beg questions about the reliability of the estimates of the degree of confidence that a particular firm is inefficient. On the other hand, given the importance of efficiency estimates in regulatory applications, an understanding of the reliability of the efficiency estimates is likely to be an important consideration. For example, it would be useful to know whether one efficiency estimate is more reliable than another. Bootstrapping would also be useful if the regulator decides the degree of inefficiency of a DMU up to a particular degree of confidence.

It is also noteworthy that in smaller samples, as the confidence intervals for efficiency estimates widen, the likely degree of upward bias in the (input-oriented) efficiency estimates also increases. Whether any correction for bias will be considered warranted will depend on how the regulator chooses to address uncertainty, as well as considerations relating to the underlying assumptions.

3.3 Returns to Scale

Different assumptions can be made in regard to returns to scale. The simplest is constant returns to scale (CRS), which essentially benchmarks relative to the highest observed productivity level (in the sense of aggregate output divided by aggregate input). This may be considered as the optimal scale from the perspective of society, being the most productive use of resources, but not necessarily the most profitable scale for the firm. Even if the technology is not CRS in general, the CRS model will be valid locally if firms are at the optimum scale (in terms of productivity).

The variable returns to scale (VRS) model in DEA imposes a constraint that the peer weights (λ's) sum to one (see discussion of the envelopment form in section 3.1). If there are varying returns to scale and firms are not all at the (socially) optimum scale, then technical efficiency estimates obtained from the CRS model will incorrectly confound scale (in)efficiency and technical (in)efficiency. The VRS model can be used to estimate the 'pure technical efficiency' for unit k (θ_k^{VRS}) because it does not include any effect of scale sub-optimality. The measure of technical efficiency of unit *k* relative to the CRS frontier (θ_k^{CRS}) is always less than or equal to θ_k^{VRS} (in the input-oriented approach), and a measure of scale efficiency (SE) can be obtained from the two as: $\theta_k^{CRS}/\theta_k^{VRS}$. In this way, DEA measures of technical efficiency under CRS can be decomposed into 'pure' (or VRS) technical efficiency and scale efficiency. When the scale of operation is not within the control of the firm, then the firm will be fully efficient if $\theta_k^{VRS} = 1$ and no inputs can be further reduced (non-radially) while still producing the same outputs (the concepts of slacks and Pareto efficiency are explained in section 3.4).

The other alternative returns to scale assumptions are non-increasing returns to scale (NIRS) and non-decreasing returns to scale (NDRS). Additionally, Petersen (1990) developed a methodology in which the assumption of convexity with respect to scale is relaxed, whilst maintaining the convexity of the input and output isoquants. This is more consistent with economic theory, since total cost functions are often assumed to be "S-shaped", which is consistent with "u-shaped" average costs.

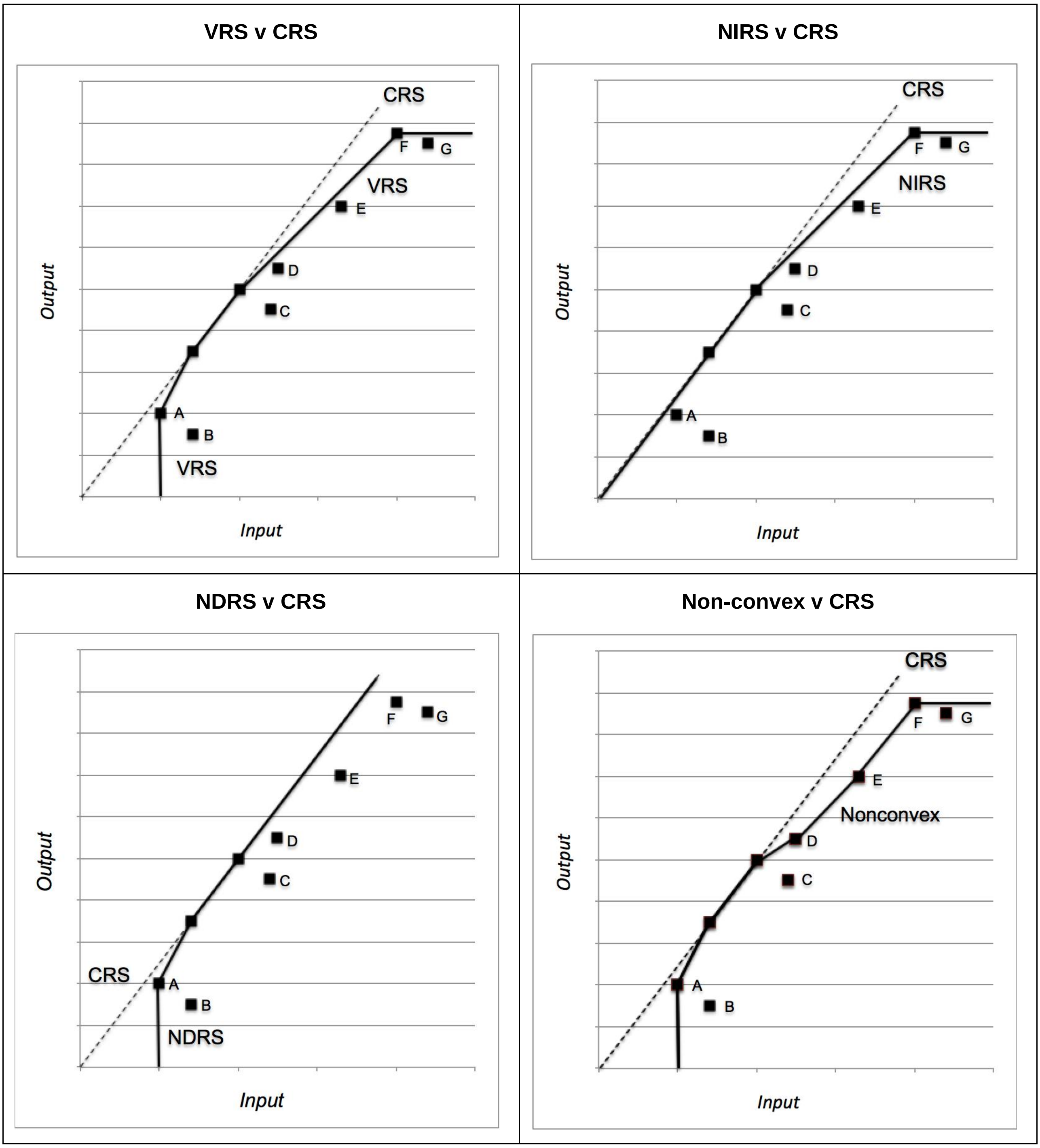
Petersen observed that convexity of the output possibility and input requirements sections is "a typical neoclassical assumption … justified by the law of diminishing marginal rates of substitution" and needed for the duality relationships to hold (Petersen, 1990, p. 307). On the other hand the assumption of a convex production possibilities set is "a restrictive assumption" which "requires marginal products to be non-increasing" which "can be relaxed under conditions of CRS and NIRS and should not be invoked under conditions of VRS" (p. 313). There is something to be said for Petersen's approach because it appears to remove an apparently unnecessary restriction in the conventional DEA model, and may therefore improve the results.

Figure 3.1 depicts the five different returns to scale assumptions mentioned. The lines represent boundaries of the feasible Production Possibilities Set (PPS) under certain returns to scale assumptions, and the area to the right of the line is the set of feasible levels of inputs that produce the corresponding output. Inefficiency is measured by the horizontal distance of a point to the frontier in proportion to the total distance to the vertical axis. Two of the observations are efficient under all of the scenarios (unlabelled) and the remaining observations are labelled from A to G. The measured degree of efficiency for these firms depends on the returns to scale assumption:

- Firm A is inefficient under CRS or NIRS but efficient under VRS, NDRS or non-convex returns to scale;
- Firm B is inefficient under all types of returns to scale, but its inefficiency is greater under CRS or NIRS than for the other scale assumptions;
- Firm C is inefficient in all cases but unaffected by the returns to scale assumption because it is projected onto a segment that is an efficient scale in all cases;
- Firms D and E are less inefficient when VRS of NIRS is assumed compared to when CRS or NDRS is assumed. However, under the non-convex model, they are fully efficient.
- Firm F is efficient under VRS, NIRS, and non-convex technologies, but inefficient under CRS and NDRS.
- Firm G is inefficient in all cases, but is more inefficient when CRS or NDRS is assumed.

It is important to recognise that firms may not be able to operate at the most efficient scale because the level of market demand is not always within their control. This is generally the case for utilities that tend to operate in discrete market areas, often as natural monopolies. Although in some circumstances there may be opportunities to modify corporate structures to alter the scale, generally we assume that the scale at which TSOs operate is essentially exogenous, and therefore scale efficiency is also outside the firm's control. This means, in general, that if an inappropriate assumption is used regarding returns to scale, the estimated technical efficiency may incorporate a scale efficiency effect that could bias the technical efficiency estimates. The issue of how to determine which assumption on returns to scale should be used is discussed in section 5.1.

Figure 3.1: **Alternative Returns to Scale Assumptions**



VRS v CRS

NIRS v CRS

NDRS v CRS

Non-convex v CRS

3.4 Efficiency Measures

3.4.1 Farrell and Pareto Efficiency

The efficiency measures obtained from standard DEA analysis represent the proportionate radial distance of the firm’s combination of inputs and outputs from the efficiency frontier, known as the Farrell measure of efficiency. If the input-oriented model is used, then the DEA-estimated efficiency score represents the proportion by which all inputs could be equiproportionately reduced (i.e. preserving the mix) while still allowing an efficient firm to produce the same output vector. Analogously, the output-oriented measure represents the maximum feasible equiproportionate expansion of all outputs using the same vector of inputs, and with the same technology.

An important limitation of the Farrell efficiency concept is that, because the mix of inputs is maintained, there may be slacks in the use of one or more inputs, or in under-producing one or more outputs. Here we focus on the input-oriented case. A ‘slack’ means that less could be used of one input whilst not using more of any other input, while still being able to produce the same output vector (with the same technology). For this reason, some of the businesses that are deemed to be efficient in the Farrell sense may not be Pareto efficient (also called Koopmans efficient). A given input vector x is Pareto efficient for producing a given output vector y if, with a reduction in *any element* of x, it would no longer be feasible to produce y (with the same technology). Hence, some firms that are Farrell efficient may not be Pareto efficient, while all Pareto efficient firms are also Farrell efficient. In other words, the set of firms that are Pareto efficient is, in general, a subset of the firms that are Farrell efficient. Pareto efficiency is the more meaningful measure for the purposes of economic regulation. Therefore, attention needs to be given to slacks.

When only a few variables are used in the DEA analysis, the number of firms with slacks may be few, but as dimensionality is increased there can be a proliferation of slacks. Since slacks are another form of inefficiency, the efficiency analysis should either account for slacks in addition to Farrell efficiency measures or use alternative efficiency measures to take slacks into account.[6] More analysis is needed to identify the Pareto efficient firms, and the methods of doing so are discussed in the next section.

3.4.2 Methods of Identifying Pareto Efficient Firms

Thanassoulis *et al* (2008) discuss a number of approaches to identify the Pareto efficient firms. The following two approaches both involve two-step procedures in which the standard DEA envelopment or multiplier program is solved as the first-stage.[7] In both cases the radial targets obtained from the first stage (i.e. each firm’s original input vector multiplied by its efficiency score) are used in place of its actual inputs in the second-stage; and in each case the second-stage uses a non-radial efficiency criterion. This is because non-radial efficiency measures take account of slacks and “have, in general, the purpose of assuring that the identified targets are on the Pareto-efficient subset of the frontier” (Thanassoulis et al., 2008, p. 268).

(a) In the first approach, the second-stage program maximises the remaining total slacks. If the optimum total slacks for the firm is zero, then it is Pareto efficient (Thanassoulis

[6] In some software the standard output reports information on slacks (e.g. the ‘dea’ user-written routine in Stata, but others do not (e.g. LIMDEP).

[7] There are also single-stage methods. Thanassoulis *et al* (p. 263) discuss one such approach designed to “arrive at once at the radial efficiency measure and at a Pareto-efficient referent point”. However, they note that “the single-stage approach may result in computational inaccuracies erroneous results” under certain circumstances. Hence, we focus here on the two-stage methods.

et al., 2008, p. 262). This method is sufficient to identify the Pareto efficient firms but does not yield a modified overall efficiency measure. However, this method may yield inappropriate results if a firm has slacks in more than one dimension, which is not uncommon (Coelli et al., 2005, p. 198).

(b) In the second approach, the second-stage program uses the input-oriented Russell efficiency measure (see: Färe and Lovell, 1978). In this measure, a given firm has a separate efficiency score for each input, and the average of those scores is the overall efficiency score for that firm. Because there is a separate efficiency score for each input, the firm is always projected onto a Pareto efficient part of the frontier. The second-stage program is applied after the radial inefficiency has been removed from the data, and finds the nearest Russell-efficient (and hence Pareto efficient) input combination in the set of feasible Farrell-efficient input combinations. This approach yields a modified efficiency measure, which is the product of the Farrell efficiency measure from the first stage, and the Russell efficiency measure from the second stage (Zieschang, 1984, p. 395). This second method has advantages over the first method since it produces a modified efficiency measure, and is not subject to the noted limitations of the first method, although it has its own limitations.

In summary, the existence of slacks is generally relevant to efficiency measurement. A firm that is found to be input-efficient may nevertheless have slacks (except in the single input case), and if so, it is not economically (i.e. Pareto) efficient. Firms of this kind would need to be identified and some account of slacks can be taken when setting the efficiency targets. The two methods discussed above can be used for identifying Pareto efficient firms, and the second of these approaches may be preferable since it yields a modified efficiency score that takes account of slacks. (A variation on this method may be to use the geometric distance function, discussed in section 3.4.3, instead of the Russell measure in the second stage.)

The identification of slacks and calculation of modified efficiency scores that take them into account does not appear to have been used in energy utility benchmarking as often as one would expect. However, for regulatory applications it may be considered important, because otherwise the efficiency of some firms may be overstated.

3.5 Output, Input and Other Orientations

Conventional DEA models have either an input- or output-orientation, and are radial, in the sense that efficiency is measured by equiproportionate contraction of inputs toward the origin, or equiproportionate expansion of outputs away from the origin (under the same technology).[8] That is, reductions of inputs, or expansions of outputs, preserve the mix. In utility regulation settings, the input-oriented measure is usually considered to be most relevant, since output is rarely a discretionary variable for these businesses, in part because regulated utilities often have an obligation to meet demand in specified locations.

However, in some circumstances it may be that both output and inputs are at least partially controllable. For example, if output quality is taken into account and quality is a discretionary

[8] The two orientations are equivalent when the technology is CRS.

variable, or if some of the inputs are either partially fixed in the short-term, or even sunk costs, then outputs and inputs may be partially discretionary. It may then be useful to consider methods that avoid the need to choose between input- or output-orientation, because they are bi-directional or non-oriented, or methods that combine elements of both orientations. A number of approaches take into account the potential for simultaneous improvement in both the input and output directions. Most of the non-oriented DEA methods also involve non-radial efficiency measures so they are each associated with an alternative efficiency measure. Such alternatives can be relevant since "under a DEA framework, no [efficiency] measure satisfies all desirable properties, so we must choose between several 'imperfect' alternatives in practice to assess technical efficiency" (Aparicio et al., 2015, p. 23).

The following are some of the non-oriented DEA methods and their associated alternative efficiency measures (see: Thanassoulis et al., 2008):

- *Additive models* involve minimising the sum of the slacks in both the input- and output-orientations. Hence inefficiency is a combination in input excess and output shortfall. One benefit of this approach is that there is no distinction between inefficiencies and slacks, as is the case in conventional DEA. Hence all benchmarks are Pareto-efficient and every inefficient firm has only one dominant peer. However, there are a number of problems and complexities in this approach, including non-uniqueness of the optimal slacks and dependence on units of measurement.
- *The hyperbolic measure of technical efficiency* involves simultaneously expanding outputs and reducing inputs by a common proportion, so that efficiency is measured against a point on the frontier between those used for the input- and output-oriented measures. Technical efficiency is measured by the maximum value of α such that $(\alpha y, x/\alpha)$ is an element of the production set.
- *The non-oriented Russell efficiency measure*: In the conventional (radial) DEA model a single scaling factor for each firm is applied to all inputs (or to all outputs). In the Russell input-oriented measure, there are separate scaling factors for each of the firm's inputs (and analogously for the output-oriented measure), and the problem is to find the optimum value of the arithmetic average of those scaling factors. In the case of the non-oriented Russell model, there are individual scaling factors for each input and output, and the technical efficiency is measured as the optimal arithmetic average of all scores: $\left(\sum_{i=1}^{m}\theta_i + \sum_{j=1}^{s} 1/\beta_j\right)/(m+s)$, where θ_i is the score for input i, β_j is the score for output j, and there are m inputs and s outputs. The 'optimum' of this average is the minimum value for which each $\left(\theta_i x_i, \beta_j y_j\right)$ is an element of the production set and $0 < \theta_i \leq 1$; $\beta_j \geq 1$.
- *Geometric distance function (GDF) efficiency measure*: Like the non-oriented Russell measures, firms have separate efficiency scores for each input and output, which ensures all sources of inefficiency are captured. However, unlike the Russell measure, this method uses geometric means of the input contraction and output expansion factors, so the problem is to solve for the minimum value of:

$$\frac{\left(\prod_{i=1}^{m}\theta_i\right)^{1/m}}{\left(\prod_{i=1}^{s}\beta_j\right)^{1/s}}$$

such that each $(\theta_1 x_1, \dots, \theta_m x_m, \beta_1 y_1, \dots, \beta_s y_s)$ is an element of the production set and $0 < \theta_i \leq 1$; $\beta_j \geq 1$. This method has the benefit that the conventional DEA radial input- and output-oriented efficiency measures, and also the hyperbolic measure above, are special cases of this model when certain restrictions are applied to the θ's and β's.

- *Directional distance functions* also use the potential to increase outputs and reduce inputs at the same time as the basis for measures of technical efficiency. The 'direction' of the distance function is determined by weights given to input reduction and output expansion, which are chosen by the analyst. This choice is arbitrary but influences the measures of efficiency obtained.

Some of these methods may have promising potential for application to TSO benchmarking if some TSO outputs are considered to be discretionary (such as quality of service, or the ability to meet peak demand) and/or if some of the inputs are considered to be non-discretionary (eg historical sunk investments that pre-date the regulatory period). The GDF measure appears to be of particular interest, because constraints can be applied to yield a variety of different models. For example, if the constraints are: $\beta_1 = \beta_2 = \cdots = \beta_s = 1$, and $\theta_1 = \theta_2 = \cdots = \theta_m = \theta$; then this represents the radial input-oriented DEA model, and the output-oriented model can be similarly imposed. More flexibly, constraints such as $\beta_h = 1$, can be imposed on selected outputs (ie exogenous outputs) while leaving some other output expansion factors to be determined subject to $\beta_k \geq 1$ (ie for discretionary outputs) and at the same time, some input contraction factors can be constrained for non-discretionary inputs, while leaving some to be determined subject to $\theta_i \leq 1$. This could potentially be a useful avenue to explore if some outputs are not considered to be exogenous and/or some inputs are not discretionary.

3.6 Controlling or Limiting Weights

As mentioned above, DEA efficiency scores are determined using a separate LP for each firm, so that in the multiplier formulation, each firm has a distinct set of (nonnegative) weights. These weights are:

> … endogenously determined shadow prices revealed by individual producers in their effort to maximize their relative efficiency. … Consequently, the range of multipliers chosen by producers might differ markedly from market prices (when they exist), or might offend expert judgement on the relative values of the variables (when market prices are missing). (Fried et al., 2008, p. 55)

The degree of variation in the resulting weights can be problematic in some applications, and it may be desirable to impose some constraints on the weights either to ensure consistency with outside sources of information, or to better reflect requirements of the decision framework. This section discusses three methods of restricting input or output weights:

(a) the general approach to restricting weights in DEA models;

(b) models that impose a common set of weights on all firms; and

(c) models in which total cost is used as a single-input, which could be viewed as imposing the constraint that the input weights be equal to input prices.

3.6.1 Subjective Weight Restrictions in the DEA Multiplier Program

The methods of imposing weight restrictions in the multiplier DEA model include using additional inequality constraints, where the boundaries of the weight restrictions are usually obtained from experts. Such restrictions can improve the reliability of efficiency comparisons where they incorporate additional information, although since they implicitly impose constraints on the technology they can detract from a key advantage of DEA (Allen et al., 1997; Thanassoulis et al., 2004). Incorporating weights can be a useful compromise between the more restrictive common weights approach and using unrestricted weights, which may be too flexible. However, it should be noted that DEA-estimated efficiency scores often show greater inefficiency when weight restrictions are added.[9] This means that care is needed to ensure that the weight restrictions are valid and do not result in underestimating efficiency scores.

Weight restrictions may be formulated in different ways. They may: (i) impose bounds on the permissible values of certain input or output weights, while allowing them to vary freely within those bounds, or (ii) they may place bounds around the ratios of different input weights or different output weights, or (iii) bounds on ratios between certain input and output weights. Restrictions of the first kind that apply directly to the values of weights are called 'absolute restrictions'. They can be problematic to formulate correctly, because the absolute values of weights do not have a clear meaning, it is the ratios of weights that have an economic interpretation. Further, they may produce unreliable results by not finding a DMU's maximum relative efficiency subject to those restrictions (Thanassoulis et al., 2008, pp. 322–323). Restrictions of the second and third kind are called 'assurance regions' (ARs), and are likely to be more reliable.

> ARs are appropriate when there is some price information and one wants to proceed from technical toward economic efficiency measures. When there is a priori information concerning marginal rates of technical substitution (transformation) between inputs (outputs) these are also suitable [weight restrictions] to use because they are based on ratios of weights that … reflect these rates. (Thanassoulis et al., 2008, p. 323)

Applications of weight restrictions to energy networks include Agrell & Bogetoft (2009, 2014) and Santos et al (2011).

3.6.2 Common Weights Models

In common weights models, the same input and output weights are applied to all DMUs. The weights are solved endogenously (not imposed) and they maximise the overall technical efficiency of the businesses being benchmarked subject to the uniformity of the weights. This

[9] This is because additional constraints in the optimization problem cannot lead to improvement in the optimal value of the objective function (i.e. the efficiency score) and lead to smaller estimated efficiencies unless the additional constraints are non-binding (or redundant).

approach produces a more stable set of weights compared to unconstrained DMU-specific weights in conventional DEA. Although in DEA each firm's weights are chosen to maximise their own efficiency, in some circumstances it may be considered more equitable for the estimated efficiencies and rankings of DMUs to be based on common weights, if that implies greater consistency of the comparisons.

The (input-oriented) common weights model can be applied by firstly solving the conventional radial DEA model for each DMU and using the computed technical efficiency scores in a further computation to find a set of common weights which minimise the squared differences between the resulting efficiency scores and the original efficiency scores (Kao and Hung, 2005).

Applications to energy networks include Saati et al (2012) and Omrani (2013). Agrell and Bogetoft (2010) discuss applications of this approach within centralised control or regulatory settings. Omrani (2013) adapted the common weights method to also take into account uncertainties in the data (i.e. possible data errors), by using a 'robust' linear programming (LP) optimisation method. In this formulation the data uncertainties are either: not stochastic, or if they are stochastic they are from an unknown probability distribution function that cannot be estimated. This methodology may be suitable if the assumptions concerning the nature of data uncertainties fit the modelling environment, and if the decision-maker wants to compare the efficiency of businesses based on common weights. Omrani applied this method to provincial Iranian gas companies.

More recently, Agrell and Bogetoft (2016a) have proposed a DEA method based on endogenous common weights (or shadow prices) which they argue is particularly well suited to normative applications such as regulation. Whereas in conventional DEA, the efficiency of each firm is maximised, in the proposed approach, a single set of weights would be determined to maximise the overall efficiency of the firms in the sample as a group. This single set of weights may then be thought of as representing the social marginal values of the inputs and outputs. The authors expect this approach to yield more stable results, in part because the more specialized or atypical firms, which conventional DEA tends to locate on the efficient frontier merely because they are in a unique part of the production possibilities set, will not necessarily be efficient using this method. The authors suggest that this would provide better efficiency incentives to the regulated firms and be more methodologically consistent.

3.6.3 Single Input (Total Cost) Models

Many benchmarking studies have used a single input total cost formulation. According to Jamasb, Pollitt and Triebs (2008) a desirable property of using total cost as a single input is that it assigns "the proper economic weighting to all inputs" (p 3402). This approach may be viewed as constraining the input weights (or shadow input prices) to be equal to the actual input prices, which may differ between each firm (depending on how total cost is calculated). In this approach, cost efficiency is equivalent to technical efficiency, based on the assumption that there is no allocative inefficiency, because the inputs are assumed to be already used in optimal proportions, and only need to be scaled back radially to achieve cost efficiency.

In some studies, the total cost measure may be constructed using standardised input prices, including the opportunity cost of capital and depreciation rates, applied to quantity data for inputs. For example, Agrell et al (2016) use an approach of this kind to measure the capital component of total costs, and they make adjustments to the labour component of non-capital cost (measured using financial data) for differences in salaries between jurisdictions. This approach to standardisation is discussed in our report 'Estimating capital costs'. This way of implementing a single-input total cost DEA model appears to be equivalent to imposing a common set of input weights equal to the standardised input prices.

Other applications of the single-input total cost model may rely more heavily on financial data to estimate total cost and convert into a common currency using either exchange rates or purchasing power parities. However, in a multilateral context there may be differences in the input prices faced by different DMUs, and this latter approach assumes that currency or purchasing parity conversion effectively deals with any differences in input price levels or relativities between the firms being benchmarked. If it does not, differences in input prices between the jurisdictions would be conflated with the cost efficiency measures, and the resulting scores could be misleading. This may be a less serious problem with the first of the two approaches described above.

3.7 Variations on DEA Methods

This section discusses latent class models, dynamic DEA, free disposal hull (FDH) and stochastic nonparametric frontiers

3.7.1 Latent class models

Conventional frontier analysis is based on the assumption that the firms in the sample have the same production technologies available to them. If there are unobserved factors that cause the production possibilities set (PPS) to differ between firms, the effects of this might be inappropriately conflated with the efficiency measures. The latent class method is designed to address the situation where some unobserved factors cause heterogeneity among the firms, and this heterogeneity is discrete, such that it divides the firms in the sample into a small number of groups (known as latent classes or subpopulations). It is often assumed that certain observable variables are related to these unobserved groups and can be used to assist in identifying them (e.g. public or privately owned firms; vertically separated, integrated or conglomerate firms; firm size; network characteristics etc.). Variables of this kind can be called 'exogenous sample separation information'.

Incorporating latent variables in DEA analysis is usually a two-step process. Prior to estimating or computing the frontier model, an analysis is undertaken to determine whether the firms fall into distinct categories, and to establish the appropriate number of classes.[10] This may be based on *a priori* choice of sample separation information, or by applying

[10] In stochastic frontier applications, a similar two-step process is often used, but there are also some one-step maximum likelihood procedures available. Some of the available one-step approaches involve using exogenous sample-separating information in the model, and others identify latent classes endogenously (see: Orea and Kumbhakar, 2004).

cluster analysis to basic descriptive variables such as productivity measures, or other statistical method (Orea and Kumbhakar, 2004). The number of groups needs to be decided and should be as small as possible, and often two classes are used. The resulting categories are summarised as latent classes. The criteria or indicators used for splitting the sample into classes should not be variables that are included in the model, because that might bias the results.

Once the set of classes is identified there are several alternative ways to proceed. In DEA applications, a separate analysis is often conducted on each class, but it is also possible to include an identifier variable within a pooled DEA analysis. A third alternative is to conduct a pooled DEA analysis, without taking into account the latent classes, and in a second-stage analysis regress the efficiency estimates on the latent variable(s) and observed operating environment variables. This latter course of action may have some benefits in the context of a small sample.[11]

Latent class models are increasingly being used in energy network efficiency analysis, including applications of SFA, DEA or both. Most studies are of the electricity distribution sector, including Culmann (2012), Agrell *et al* (2013), Dai and Kuosmanen (2014), and Orea and Jamasb (2014). Culmann tested firm size as a differentiating characteristic for German electricity distributors and, using a 'one-step' SFA latent class technique, found that larger businesses operate under a different technology than the smaller businesses. Agrell *et al* used a latent class regression analysis as an initial step to identify four separate groups of Norwegian electricity distributors, based on network characteristics, and subsequently carried out various benchmarking methods separately on the four sub-samples. In a study of Finnish electricity distribution, Dai and Kuosmanen proceed in the opposite direction, firstly estimating a benchmarking model (using the 'StoNED' semi-parametric regression method), and then using the results in a cluster analysis to define groups within the benchmarked firms. In particular, certain clustering criteria were calculated from the efficiency scores (namely the ratios of each of the output quantities against the estimated efficient cost for each DMU) and the normal mixture model (NMM) clustering technique was employed. Dai and Kuosmanen distinguish between absolute benchmarks (fully efficient firms) and relative benchmarks (the most efficient firms in each cluster), which may not be fully efficient. Orea and Jamasb analyse Norwegian electricity distributors using a latent class SFA approach combined with the recently developed 'zero inefficiency SFA' model, and find some distinct differences among utilities related to different weather conditions and locations. In an application to electricity transmission, Llorca at al (2014) examined the cost efficiency of 59 TSOs in the USA over the period 2001-2009. The study found that the average efficiency score increased substantially in moving from one to two classes, but there were only incremental changes to the average efficiency score when further increasing the number of classes.

Latent class analysis may be useful if it is believed that there are unobserved characteristics that cause systematic differences in the technological possibilities of firms within the sample. However, the literature in this field is relatively recent which suggests that these methods are

[11] In econometric methods such as SFA, it is possible to pool the model and allow some subset of the parameters to vary between the classes.

yet to be thoroughly explored. To our knowledge, no regulatory authority has used this approach in a regulatory setting.

Moreover, limitations of sample size may diminish the opportunities to use this approach to benchmarking European TSOs if it means conducting separate analyses on sub-samples of data. Although the statistical significance of the differences between the subpopulations could in principle be tested, this may not be informative in small samples. It should also be noted that dividing a dataset into subsets would almost invariably cause DEA efficiency scores to increase (due to increased estimation bias) as the sample size decreases.

As an alternative approach, it may be feasible to use latent variables together with observed operating environment variables in a second-stage analysis to control for, and then eliminate, the effects of unobserved differentiating factors from the efficiency scores. Issues of interpretation of the latent classes and the reliability of the latent variables as indicators of unobserved factors can be addressed at that later stage of the analysis.

3.7.2 Dynamic DEA

This section briefly discusses the approach known as dynamic DEA. These methods employ DEA in a multi-period setting and seek to identify the degree of dynamic efficiency. The optimisation problem spans over a series of periods because one or more of the inputs is ‘quasi-fixed’, in the sense that it cannot be adjusted to the optimal level instantaneously, but in a later period.

Von Geymueller (2009) employed a dynamic DEA model to study the dynamic efficiency of 50 electricity TSOs in the USA from 2000 to 2006, and the quasi-fixed inputs were transmission lines and transformer stations. In the first year of the sample period, the inputs included the variable inputs and the quasi-fixed inputs carried over from the previous period. The produced outputs in the first period were the quantity of services supplied to customers (energy delivered) and the quantities of quasi-fixed inputs at the end of the period. That is, investment (or disinvestment) activity was treated as part of the output of the firm, and the quasi-fixed assets created in the first period became inputs in the next period. Because the quasi-fixed input is both an input (at the beginning of the period) and an output (at the end of the period), a non-oriented approach is needed, and von Geymueller used the additive model. Although annual data for variable inputs and service outputs is used in the analysis, only the starting value of the quasi-fixed input is used (other values being those obtained endogenously as the solution or optimal values at the end of each period).

Von Geymueller found the dynamic efficiencies of every firm were higher than the static efficiencies (obtained from the static version of their production model), which seems to be a general result. The static model suggested the TSOs “were persistently over-equipped with quasi-fixed inputs”, whereas the dynamic model indicated that TSOs were increasingly under-equipped with quasi-fixed inputs, which explained the efficiency deterioration over the period found with both static and dynamic measures of efficiency. Von Geymueller concluded that in the case of TSOs, where provision of services depends crucially on investment in long-lived assets, static measures of efficiency may misleadingly understate the degree of efficiency, and he suggested that regulators “should definitely have a look at dynamic efficiencies and not rely on a static efficiency analysis only” (von Geymueller, 2009, p. 412).

3.7.3 Free Disposal Hull (FDH)

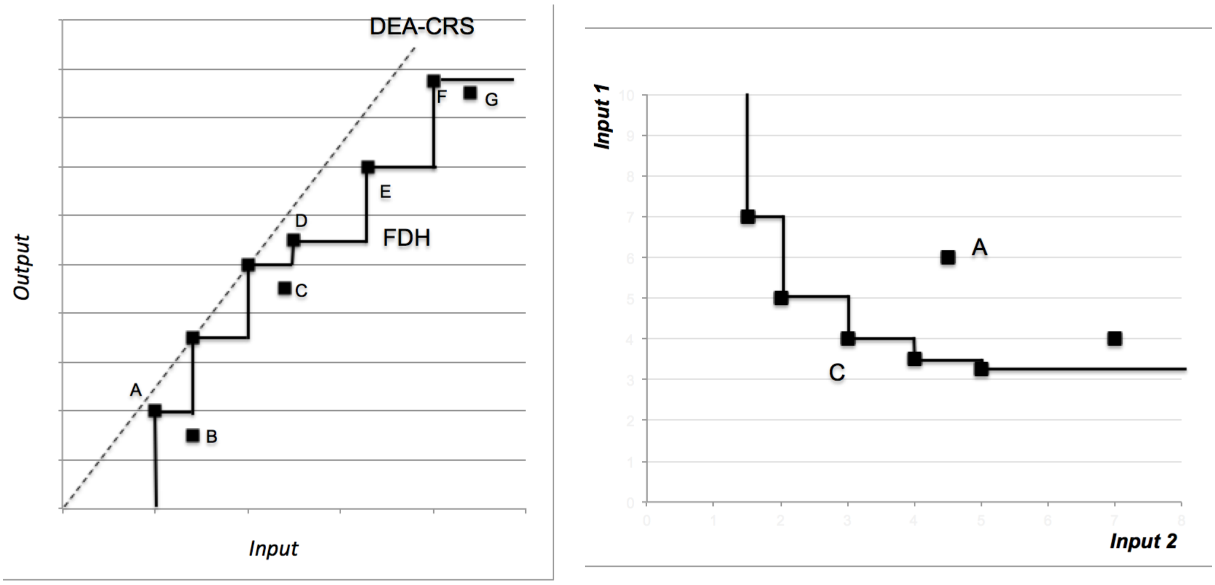
The Free Disposal Hull (FDH) method is an alternative nonparametric frontier method in which the input and output requirement sets are not constrained to be convex, and only the assumption of free disposability of inputs and outputs is relied on. Like DEA, the efficient firms determine the position of the frontier, however, the frontier is not characterised by convex combinations of those points (or conical for CRS), as with DEA. Instead the frontier extends down vertically and to the right horizontally from each of these points.[12] Figure 3.3 shows the FDH with a single input and single output dimension. In this approach, inefficient firms are usually projected onto an efficient (dominant) firm, after removing slacks, rather than onto a weighted average of efficient (peer) firms. "It needs to be emphasized that the principal merit of FDH analysis is that it always uses a single actually observed input-output bundle as the basis for comparison and efficiency evaluation of any firm" (Ray, 2004, p. 139).

As previously mentioned, the convexity of input requirement and output possibilities sets are usually viewed as standard assumptions in neoclassical economics, so in general DEA can be viewed as better conforming to economic theory than FDH. However, the FDH method may be useful if there are strict limitations on input substitution such that inputs can only be used in a limited number of combinations. Daraio and Simar suggest there are other circumstances where this method may be relevant. "FDH technical efficiency measures remain meaningful for theories of the firm that do allow for imperfect competition or uncertainty" (Daraio and Simar, 2007, p. 38).

There do not appear to be many applications of FDH in energy network benchmarking. Perhaps its most useful feature is the identification of the dominant firm corresponding to each inefficient firm. Dominant firms are a subset of the peers identified in DEA, and are the most important peer for an inefficient firm to have regard to. The FDH may therefore be useful as a supplementary exercise to check sensitivity to the convexity assumption and to identify the dominant peers.

[12] The mathematical formulation for FDH is the same as for DEA-VRA except that the λ's are constrained to be either 0 or 1 (rather than simply ≥ 0).

Figure 3.3: **Free-Disposal Hull Frontiers**



3.7.4 Stochastic Nonparametric Frontiers

If the data is considered to be subject to inaccuracies and/or outliers that can be treated as random noise, then the constraint that all observations must be on one side of the frontier can become a significant limitation, and in these circumstances various methods can be used. In DEA, outliers are normally removed, and the identification of outliers is discussed in section 5.3. This section discusses alternative nonparametric methods that directly take noisy data into account. Parametric methods such as stochastic frontier analysis (SFA) are outside the scope of this paper, but are briefly discussed in sections 6.1.2 (on comparing DEA results to parametric and index methods) and chapter 8 (on combining models).

One stochastic nonparametric approach is ‘chance constrained programming’ (Land et al., 1993; Olesen and Petersen, 1995). Explicit assumptions are made about the distributions of the random element of the sample data, and the constraints are expressed in probabilistic terms. Information on the parameters of the probability distribution is also needed, and overall the “data requirements of chance constrained programming are severe” (Fried et al., 2008, p. 58).

Partial frontier approaches (order-m and order-α frontiers) generalize FDH by combining it with a probabilistic approach (Tauchmann, 2011). They involve locating a certain quantity or percentage of super-efficient observations (i.e. those beyond the estimated efficiency frontier), and can thereby reduce the sensitivity of the estimated frontier to outliers. Hence, for an extreme observation, *j*, that is always on the frontier when included (and hence $\theta_j = 1$), when excluded will have $\theta_j > 1$ (in the input-oriented case), and if its average score over the simulations is greater than 1 it is considered to be ‘super-efficient’. The frontier estimated in this way will be robust to outliers, although there may be limitations as a basis for ranking efficiencies (and dependency on the choice of m or α).

Another set of approaches is based on nonparametric regression, which estimates mean values of a dependent variable based on a given set of covariates without specifying any functional form (see: Parmeter and Racine, 2013; Kuosmanen et al., 2015; Kumbhakar et al., 2017). For example, the model may be of the form: $y = g(x_1, x_2, x_3) + \epsilon$, where ϵ is a random noise term and no functional form is specified for $g(.)$. It is estimated using nonparametric methods (e.g. kernel based local least squares or local likelihood). Models of this kind may be extended to incorporate separate distributions for the inefficiencies and the noise, and various stochastic assumptions. These techniques are relatively recent but have found application in utility regulation.

One approach of this general class is Stochastic Nonparametric Envelopment of Data (StoNED), which seeks to encompass both stochastic frontier analysis and DEA within a more general framework, and was adopted by the regulatory authorities in Finland in 2012 (Kuosmanen et al., 2015). The StoNED method has two steps. The first step is to estimate a cost function using convex nonparametric least squares (CNLS) regression. The cost function has a specific form in which the log of total cost is a linear function of: the log of expected total cost; the operating environment effects; and the conflated stochastic term. The expected total cost is a linear function of the outputs, where the coefficients are specific to each firm and can be interpreted as marginal costs. In the second stage of the analysis, estimates of technical efficiency are obtained from the conditional expectations of the inefficiency component of the half-normal error term, which is analogous to the SFA model (Kuosmanen, 2012).

In the current StoNED benchmarking method used for distribution networks in Finland there are two inputs: (a) controllable operating costs; and (b) the replacement value of the network. These are not combined in the efficiency analysis, which only applies to opex, whereas the value of the network is treated as a fixed input with no efficiency target (Kuosmanen, 2012, p.81). The opex benchmarking analysis uses the StoNED method, which is a form of non-parametric regression, which takes the following form:

$$\ln x_1 = \ln G(x_2, y) + \delta' z + u + v$$

where:

- x_1 is (deflated) opex, x_2 is the (deflated) asset replacement cost, and y is a vector of outputs.
- $G(.)$ is the nonparametric 'input needs function'.
- z is a vector of operating environment factors that cause heterogeneity and enter the model linearly and not as part of the nonparametric 'input need function'. The δ's appear to be estimated in a similar way to regression parameters.
- v is white noise and, similar to SFA, u is a one-sided stochastic term that measures inefficiency, but "is estimated without distribution assumptions using the nonparametric kernel deconvolution method" (p.83).

There are four outputs: (i) the volume of transmitted energy (GWh); (ii) the length of the electricity network (km); (iii) the number of metering points; and (iv) regulatory outage costs

(EUR)—a ‘bad’ output. The operating environment variable is the ratio of connections to metering points.

It should be noted that there are about 80 electricity distribution businesses in Finland and the data covers a number of years, so that there is a substantial amount of data used in the nonparametric regression analysis. It is unlikely that this method could be reliable with substantially smaller datasets.

3.8 Operating Environment Variables

There are several ways of taking operating environment variables into account (see: Coelli et al., 2005, pp. 190–195). This section briefly discusses three of them.

One approach is to treat the environmental variables as additional inputs or outputs in the DEA analysis. For example, Frontier et al (2013) appears to have taken this approach by including population density as a cost driver. This may be contentious because efficiency measurement in DEA assumes that the outputs are produced from the inputs. There is no reason to expect that assumptions derived from production theory, such as convexity, would apply if this were not the case. Furthermore, including operating environment factors in the DEA analysis as if they were inputs or outputs inevitably has the effect of making more units appear efficient.

A second approach also involves including the operating environment variables in the DEA envelopment program, but treated separately from the other variables by including a separate constraint relating to the operating environment variables (or two constraints if some of the operating environment variables enhance output while others retard output). This approach has three notable drawbacks: (a) it is necessary to know in advance whether an operating environment variable increases or reduces output; and (b) the operating environment variables don’t influence the efficiency with which a firm uses inputs to produce outputs; and (c) the operating environment variables must be continuous (they cannot be categorical variables).

The third approach involves firstly carrying out the DEA analysis without controlling for the exogenous factors, and then conducting a second-stage analysis, in which the estimated efficiencies are used as the dependent variable in a regression against the operating environment factors. A Tobit (i.e. censored) regression model has most commonly been used in the second-stage to take account of the fact that the maximum efficiency score is one. More recently, Simar and Wilson (2007) have questioned the validity of censored regression, and instead advocated truncated regression. They developed single and double bootstrapping procedures that enable valid inferences from the second-stage regression (under certain statistical regularity conditions). In the estimated second-stage model, the signs of the coefficients on the operating environment variables indicate the directions of their effects and “standard hypothesis tests can be used to assess the strength of the relationships” (Coelli et al., 2005, p. 194). Importantly, the model obtained from the second stage regression can be used to calculate adjusted efficiency scores, which control for differences in the operating environment factors (by substituting sample means for the environmental factors).

According to Fried et al (2008), the two-stage approach is by far the most popular of these alternatives, and Coelli *et al* “recommend the two-stage approach in most cases” (2005, p. 194). This approach has several advantages:

- It can accommodate categorical as well as continuous environmental variables;
- It can accommodate more environmental variables than could be incorporated within the DEA program;
- It does not make prior assumptions about the direction of the influence of the environmental variable;
- Hypothesis tests can be used to test the significance of the influence of an environmental variable in the efficiencies;
- It is simple and transparent.

Second-stage regression of efficiency scores on operating environment factors is discussed in more detail in section 6.5.

4 REGULATORY BENCHMARKING METHODS

This section reviews the methods of regulating electricity and gas TSOs used in a number of countries and the benchmarking methods and applications used. In 2001 Jamasb and Pollitt (2001) found that only two regulators (the Netherlands and Norway) had undertaken significant electricity TSO benchmarking. In 2012, Haney and Pollitt (2012) reported that 13 out of 25 regulators surveyed used some form of economic benchmarking for electricity TSO regulation. Only four used frontier methods (the Netherlands, Finland, Portugal and Brazil) with most of the remainder using unit cost benchmarking (often for specific types of costs or activities) or reference network analysis (or both). Five of the 12 regulators that did not use benchmarking were giving consideration to using it. These observations tend to suggest that the use of benchmarking in TSO regulation is increasing over time.

This chapter discusses the methods of TSO price regulation used by a number of regulators, including their use of benchmarking where applicable. This survey is limited in scope and aims to provide no more than an overview of the roles of benchmarking within international energy TSO regulation.

4.1 Japan

Japan has primarily vertically integrated electricity and gas businesses, and rates are regulated from time-to-time when a business applies for a tariff increase for small users. A yardstick regulation framework is used in which costs are assessed by benchmarking direct costs (in two broad categories) using regression analysis of cost drivers. Businesses are grouped into regions and into three broad efficiency categories, and the allowed tariff changes depend on the efficiency category the firm is in (ACCC and AER, 2013).

4.2 Finland

Finland has approximately 80 electricity distribution networks, 12 regional high-voltage distribution network operators, and one electricity TSO (Fingrid) which is majority owned by the state. The gas industry is an integrated statutory monopoly and not discussed here.

The Energy Authority is the relevant regulator, and the form of regulation applying to each of these three types of electricity networks is different. For the electricity distribution networks, the Energy Authority applies a revenue cap with surpluses or deficits calculated annually. The revenue cap is based on an assessment of reasonable and efficient costs plus a rate of return on invested capital, the rate of return being based on the capital asset pricing model. Benchmarking is also used to set company-specific efficiency targets (see section 3.8.2). For the high voltage distribution networks the regulatory formula for efficient costs is much simpler, with just a general efficiency target (in per cent) for these businesses.

In the framework applying to the electricity TSO, Fingrid can set its own prices subject to them being equitable and non-discriminatory, and reasonable as a whole. The Energy Authority makes an *ex ante* assessment of compliance with these requirements over a four-year period, using a rate of return or "cost-plus" method. It also uses negotiation supported by benchmarking of totex, using several benchmarking methods, including DEA, SFA, COLS and unit costs, to make its assessment (Haney and Pollitt, 2012).

4.3 Germany

Germany has four major electricity TSOs which are also major generators. There are over 800 local electricity distribution companies. The *Bundesnetzagentur* (BNetzA) is responsible for regulating prices and competition in utility sectors.

In 2009, incentive regulation was introduced to encourage investment in energy-related infrastructure. A similar form of regulation applies to the electricity and gas sectors, and to transmission and distribution businesses (except for small electricity distributors with less than 30,000 customers). Under this framework, cost efficiency benchmarks are used to determine *ex ante* the allowed revenue and approved investment budgets for each regulatory period. Network operators are obliged to benchmark their controllable costs against the costs of other network operators with a similar structure. The resulting estimates of efficient cost form the basis of realistic efficiency targets, and any prevailing cost inefficiencies are required to be removed over two five-year regulatory periods. The efficiency targets also include industry-wide productivity improvements due to technical change.

A number of detailed aspects of the benchmarking framework are stipulated in legislative instruments. For example, cost drivers must include connections, circuit length and peak load for electricity networks. Two modelling methods are to be used: DEA with non-decreasing returns to scale (NDRS) and SFA, and two measures of capital inputs are to be used, book value and standardised capital cost. This results in four estimates of efficiency for each firm. The highest estimate is chosen, or 0.6 if that is higher (Agrell and Bogetoft, 2016). Other analysis was undertaken by BNetzA to inform its efficiency benchmarking analysis, including engineering-based reference network modelling to identify possible cost drivers.

4.4 Ireland

Ireland has a vertically integrated electricity industry, with the exception of the retail market, which is open to competition. The electricity transmission and distribution networks are owned and operated by the Electricity Supply Board (ESB), which is predominantly state-owned. Ireland's gas transmission and distribution networks are owned by Bord Gais Eireann (BGE), a government-owned entity, and its subsidiary Gaslink is the independent system operator. The Commission for Energy Regulation (CER) is responsible for electricity and gas market regulation.

The CER determines the revenues that transmission businesses can earn over a five-year period and then refines the revenue cap annually. To support this role CER carries out totex benchmarking using unit costs and reference network analysis (Haney and Pollitt, 2012).

4.5 Netherlands

The Dutch gas network consists of separate networks to transport low calorific gas (used by small users) and high calorific gas (used by industry and power generators). There is one national TSO for gas in the Netherlands, Gasunie Transport Services (GTS), which is fully state-owned, and there is one electricity transmission system operator, Tennet, which is also state-owned. There are eight DSOs that distribute both gas and electricity and two DSOs that distribute gas only. These are owned by municipalities. The regulator is *Autoriteit Consument*

& Markt (ACM), which is responsible for both competition law enforcement and utility regulation.

The electricity and gas DSOs are subject to price regulation with a system of national yardstick competition. The TSOs are subject to revenue caps with a yardstick that is partly based on international benchmarks.

4.6 Portugal

The Portuguese electricity transmission industry is vertically separated with a number of interconnections with the Spanish grid, and both nations operate a single electricity spot market for the Iberian Peninsula. REN is Portugal's only electricity TSO, operating the transmission network under a 50-year concession. REN is a formerly state-owned and now privatised entity. The Energy Services Regulatory Authority (*Entidade Reguladora dos Serviços Energéticos*) (ERSE) is responsible for regulating the electricity market. ESRE negotiates electricity transmission charges with REN and uses benchmarking of total expenditure to inform those negotiations. The benchmarking methods include DEA, corrected ordinary least squares (COLS), SFA and reference network analysis, applied to an international sample of electricity TSOs (Haney and Pollitt, 2012).

REN is also the only gas TSO in Portugal, operating the natural gas transmission grid under a 40-year concession agreement, including underground storage facilities and LNG terminal. ERSE regulates gas transmission access tariffs.

4.7 Sweden

Energy networks in Sweden are regulated natural monopolies. The National Electricity Grid is operated by a single state-owned TSO, Svenska Kraftnät. There are five companies that operate regional electricity networks, and 73 that operate local electricity networks, with a range of ownership arrangements including state, municipal, private and other. The Swedish gas transmission system is owned and operated by Swedegas, which is owned by the Spanish and Belgian gas network companies Enagás and Fluxys.

The Swedish energy regulator is the Energy Markets Inspectorate. Until about 2012, the Inspectorate carried out annual reviews of energy network tariffs using a reference network performance assessment model (NPAM), which estimated a monetary value of the services provided by the utility to its customers (taking into account outages etc), which could then be compared to the amounts it had charged.

The NPAM was replaced by *ex ante* revenue cap regulation in which the Inspectorate receives pricing proposals from businesses and decides on a revenue cap, usually for a four-year period. For TSOs', benchmarking was not used in this assessment, and instead a relatively arbitrary efficiency improvement factor such as 1% per year was imposed.

4.8 United Kingdom

Great Britain has three electricity TSOs (National Grid electricity transmission, Scottish Hydroelectric Transmission Limited, and Scottish Power Transmission Ltd) and one gas TSO

(National Grid gas). There are 14 regional electricity distribution networks (owned by seven firms) and four gas distribution networks.

The energy regulator is the Office of Gas and Electricity Markets (Ofgem). It administers an RPI – X price cap regime with regulatory periods of eight years now known as the ‘Revenue using Incentives to deliver Innovation and Outputs’ (RIIO). The same regulatory approach is used for TSOs and DSOs, electricity and gas. Within this approach, Ofgem estimates the expected efficient total expenditure (‘totex’ = opex + capex) using a number of different methods. In its November 2014 decision for electricity distribution networks, the following three cost assessment methods were used:

- Benchmarking using regression analysis of totex against two main cost drivers: a composite scale variable (CSV) of modern equivalent asset value; and customer numbers.
- Examination of disaggregated activity costs using various cost assessment techniques such as regression analysis, ratio analysis, trend analysis and technical assessment (the methods varying between activities). The assessed efficient activity costs were then aggregated to obtain a ‘bottom up’ estimate of efficient totex. The cost drivers for each activity in the ‘bottom-up’ analysis were also aggregated to obtain a different a different CSV measure.
- A second benchmarking analysis in which totex was regressed against the second CSV measure.

Ofgem then combined these three models giving 25% weight to each of the two regression models and 50% weight to the disaggregated analysis, to obtain Ofgem’s estimate of expected efficient totex. Then the upper quartile estimate of efficient totex was estimated for each of the three modelling approaches, and these were combined using the same method to obtain Ofgem’s estimate of the upper quartile totex. Finally, Ofgem assigned 75% weight to this upper quartile estimate and 25% weight to the regulated business’ own cost forecast.

4.9 Brazil

There are approximately 13 major electricity transmission businesses in Brazil. They are for the most part government-owned (either at federal or state level). The number of TSOs is rapidly increasing due to recent concession auctions for around 40 transmission networks constructed since 2000. The Brazilian electricity regulator (ANEEL) has benchmarked electricity TSOs since 2007 using DEA. In the models used, the input is operating costs, typical outputs are the numbers of power transformers and switch modes, transformer capacity and network length, and returns to scale are assumed to be non-decreasing. In 2013, ANEEL used panel data for nine TSOs (da Silva et al., 2017). More comparators will be available for future benchmarking exercises.

In the natural gas industry, state-owned Petrobrás is a major vertically integrated gas producer and owner of transmission and distribution networks. It controls more than half the combined capacity of gas transmission pipelines and has a controlling interest in the majority of gas distribution networks. The natural gas sector in Brazil is regulated by ANP (the National Agency of Petroleum, National Gas and Biofuels). ANP’s regulatory role was strengthened in 2009, with responsibility for approving transmission tariffs and access agreements for gas pipelines.

4.10 New Zealand

There are several gas TSOs in New Zealand, including Powerco, Vector and GasNet, with varied forms of ownership including private, municipal and community trusts. Natural gas TSOs have been subject to price regulation by the New Zealand Commerce Commission (NZCC) since 2012. The form of price control is a revenue cap, with price paths initially based on pre-existing prices and escalation (in real terms) based on long-run average productivity changes for gas networks relative to the economy as a whole. These were estimated using a TFP index methodology (Economic Insights, 2011; ACCC and AER, 2012).

The only electricity TSO in New Zealand is state-owned Transpower, which is also regulated by the NZCC since 2011. The NZCC sets an individual price-quality path for four and five yearly periods, which specifies the maximum allowable revenue, expenditure allowances, and required quality standards. The revenue cap was derived using a building block methodology.

4.11 USA

In the USA, the Federal Energy Regulatory Commission (FERC) regulates interstate gas transmission. The 30 largest pipeline companies operate the vast majority of interstate gas transmission pipelines. Most gas TSOs are vertically separated from other functions. Although the gas transmission market is reasonably competitive in North America, FERC must approve any increases in tariffs for individual pipelines to ensure they are ‘just and reasonable’. Approved tariffs effectively become a price cap until another rate case is held. The FERC uses a cost of service approach, which includes quantifying reasonable operating and maintenance expenses and the allowed return on capital invested and used to serve customers.

The FERC is also the regulator of interstate electricity transmission, and this covers most of the USA’s highly interconnected grid. Electricity TSOs are mostly privately owned, although many are owned by (non-profit) cooperatives. Often they are vertically integrated, but are required to be functionally unbundled. Each electricity TSO is required to have standard open access terms and conditions and can submit tariffs to FERC for approval. The FERC uses the same cost of service for determining electricity transmission rates.

4.12 Australia

There are five electricity TSOs in Australia’s national electricity market (which excludes Western Australia (WA) and the Northern Territory), all of which are regulated by the Australian Energy Regulator (AER). Three are privately owned and two state-owned. When determining electricity TSO tariffs the AER uses a ‘building block’ approach which involves forecasting the cost of supply (including a commercial rate of return on assets) and deriving revenue caps consistent with those costs. As part of this exercise it must assess whether the expenditure projections submitted by a TSO is consistent with criteria such as efficiency.

This depends on the efficiency of past expenditure and on the reasonableness of the projection from past to future expenditure, including the forecasts for outputs. The AER aims to use economic benchmarking and category analysis (i.e. unit cost benchmarking by cost category), not only to assess the efficiency of past expenditure, but also as one of the methods used for assessing the reasonableness of projections. The methods it is developing include multilateral total factor productivity, data envelopment analysis and econometric modelling (AER, 2013).

The AER developed its approach to benchmarking electricity TSOs in 2013 and 2014, and has presented benchmarking results for the five TSOs in three annual benchmarking reports since then. Multilateral total factor productivity and multilateral partial factor productivity indexes are used to measure the relative productivity of TSOs and productivity changes over time. The AER currently uses economic benchmarking in its price decisions to derive its forecast of future productivity changes used in assessing TSO opex forecasts. It does not currently use benchmarking to make efficiency adjustments for particular (inefficient) TSOs. This is due to a lack of consensus on measurement of outputs for transmission networks. The AER is currently carrying out a review of its transmission benchmarking methods.[13]

Gas transmission pipelines in Australia (excluding those in WA) are potentially subject to regulation by the AER. The access regulation framework only applies if a pipeline meets certain declaration criteria (including that their use is essential to supply a significant downstream market) or do not benefit from a 15-year ‘no coverage’ period. Out of 14 major gas pipelines in Eastern Australia, seven are subject to access regulation and seven are uncovered. These pipelines are all privately owned. The APA Group is the largest owner of gas pipelines, and it owns all of the seven declared pipelines, and has ownership interests in two of the uncovered major pipelines. The next largest pipeline owner is a consortium of Jemena and Singapore Power, which owns three major uncovered pipelines.

There are two different forms of regulation, named ‘full’ and ‘light’ access regulation.[14] Both require the gas TSO to publish standard terms and conditions for access. However, with ‘full’ regulation, the terms and conditions, including the tariffs, are subject to approval by the AER, which is not the case with ‘light’ regulation. Instead, the AER monitors tariffs and can arbitrate disputes between the TSO and customers. Light regulation is used where the costs of full regulation are considered to be disproportionate to the benefits.

When determining pipeline tariffs under ‘full’ regulation the AER assesses the cost of supply (including a commercial rate of return) using a ‘building block’ method and derives reference tariffs consistent with that cost. The AER has not yet used benchmarking in its regulatory decisions for gas transmission pipelines. Nor has it used productivity analysis for the purpose of setting the ‘x-factors’. In a recent draft decision the AER said that, while it usually forecasts productivity growth “based on historic industry productivity performance as measured by econometric modelling”, it does not yet have an adequate dataset for gas transmission to enable modelling of that kind (AER, 2017, pp. 7–14).

[13] https://www.aer.gov.au/communication/aer-invites-submissions-on-review-of-transmission-benchmarking

[14] Four of the seven regulated pipelines are under ‘full’ regulation and three are under ‘light’ regulation.

4.13 Conclusions

Most of the regulators of gas and electricity TSOs discussed in this chapter use some form of benchmarking as part of the analysis they undertake. Various benchmarking methods are used, for the most part unique to each regulator. Use of benchmarking is more widespread among the European regulators discussed here compared to those outside Europe. In Finland, Germany, the Netherlands and Portugal, benchmarking techniques such as DEA, COLS and/or SFA are used and benchmarking has a key role in the regulatory frameworks applying to most TSOs. Ireland and the UK appear to use more simplified cost benchmarking methods, while Sweden does not appear to use benchmarking for TSOs. Outside Europe, countries in which benchmarking has a key regulatory role include Brazil and Japan. In the USA, the regulator, FERC, does not appear to make substantial use of benchmarking for electricity and gas transmission. In Australia and New Zealand, benchmarking is not as developed for TSO regulation, although these countries have made considerable use of benchmarking for energy distribution networks. In Australia the AER is currently developing its benchmarking method for electricity TSOs, but it is not clear whether benchmarking will be extended to gas TSOs.

5 SELECTING A PREFERRED MODEL

This chapter examines several topics related to model selection. Although the topics addressed in chapters 6 and 7 logically follow after the selection and estimation of a model, because benchmarking is generally an iterative process, they are also relevant to model selection.

The model selection process in DEA has parallels to model selection in regression. When DEA is viewed as a non-statistical model, it is difficult to establish analytically whether a particular model adequately represents the 'true' underlying efficiency frontier (Greene, 2007, pp. E33-110). This involves questions about whether the ranking of the efficiency of utilities is accurate and whether the efficiency estimates themselves are sufficiently robust. The recent developments in statistical foundations for DEA (including bootstrapping) address this issue.

5.1 Returns to Scale

At the outset of a benchmarking exercise it is necessary to decide which assumption about returns to scale is most plausible for the industry at hand. For example, whether the technology is constant returns to scale (CRS), variable returns to scale (VRS), non-increasing returns to scale (NIRS) or non-decreasing returns to scale (NDRS). As discussed in section 3.3, this choice is important in DEA, because it influences the estimated measures of efficiency, and if the wrong assumption is made the estimates will generally be inconsistent. Badunenko and Mozharovskyi (2016) recommend that "[b]ecause of the importance of the returns to scale assumption for the DEA estimator, this data-driven test should be performed before applying any DEA model."

The returns to scale assumption can be based on expert knowledge about the industry. However, if not certain, then scale economies may be tested using an econometric model, or alternatively in the DEA context, Simar and Wilson (2002) have developed tests that can be employed, based on scale efficiency measures and using bootstrap methods. Nonparametric approaches of this kind remain at an early stage of development, require large data samples, and are not yet widely used. Preliminary econometric estimation of the cost, distance or production function using SFA is the most common way of ascertaining the nature of scale economies.

Since the main focus of this paper is nonparametric methods, the Simar and Wilson methodology is briefly outlined. This method based on measures of scale efficiency, which are defined in terms of distance functions measured under different assumptions about returns to scale. Distance functions are normalised measures of the distance of an observation (x, y) from the frontier (i.e. the boundary of the production set). For example, in Figure 3.1, the input-oriented distance function for an observation is the horizontal distance of that observation to the vertical axis, divided by the horizontal distance to the vertical axis of the point on the frontier that it is projected onto. The distance function is the reciprocal of the Farrell efficiency measure: $D(x, y) = 1/\theta \geq 1$.

The measures of scale efficiency are: $s(x, y) = D^{crs}(x, y)/D^{vrs}(x, y)$; and $g(x, y) = D^{nirs}(x, y)/D^{vrs}(x, y)$, where the superscript crs refers to constant returns to scale (CRS);

vrs refers to variable returns to scale (VRS); and *nirs* refers to non-increasing returns to scale (NIRS). If $s(x,y)=1$ there are local CRS for the observation (x,y). If $s(x,y)>1$ and $g(x,y)=1$ there are local decreasing returns to scale, and if $s(x,y)>1$ and $g(x,y)>1$ there are local increasing returns to scale. These relationships form the basis of statistical tests in the Simar and Wilson methodology. The test statistics are based on $s(x,y)$ and $g(x,y)$ above, but aggregated over all firms in the sample, and bootstrapping is used to derive the probability distributions, and hence the critical values for the test statistics.

The aim of the tests is not to determine the nature of *local* returns to scale, but to determine the nature of the *global* returns to scale (i.e. over the full range of feasible values of inputs and outputs). The process involves firstly testing the null hypothesis that the frontier is globally CRS. If that hypothesis is rejected, then the second test is carried out, which is to test the null hypothesis that the frontier is globally NIRS. If that hypothesis is also rejected then it is concluded that the frontier is VRS. Simar and Wilson (2002) carried out Monte Carlo tests to validate these tests using 2 or 3 variables and sample sizes of 20, 40 and 60, and found them to perform satisfactorily. A very recently published alternative DEA-based approach for testing hypotheses, including returns to scale, is Kniep et al (2016).

5.2 Choosing the final outputs and inputs

One of the questions to be addressed is whether there is any means of establishing the most appropriate configuration of inputs and outputs. Unnecessary variables need to be excluded otherwise they increase the number of DMUs found to be efficient, often spuriously, leading to an exaggeration of efficiency estimates. Highly collinear variables can lead to distorted efficiency estimates, which may require aggregating variables. Simar and Wilson have observed that in DEA generally, “it is very important to eliminate any inputs or outputs that are not truly part of the production process” (2001, p. 167). It is equally important to “exploit any opportunities for aggregation that might exist” (Simar and Wilson, 2001, p. 161). This may be especially important given the small data samples used for TSO benchmarking. That said, just as adding variables to the DEA model results in more units appearing to be efficient, eliminating or aggregating variables tends to lower the average efficiency estimates, except in special cases. This is why formal tests for variable exclusion or aggregation can be particularly useful. This section describes two of the approaches that have been commonly used in practice, and briefly recaps a number of other methods discussed in our report ‘Selecting cost drivers’.

5.2.1 Stepwise DEA

Stepwise DEA methods are an adaptation of stepwise regression to a non-parametric setting. The aim is to test the significance of changing the variables by adding additional variables or by replacing an aggregated variable with its disaggregated components.

Kittelsen (1993) developed a stepwise DEA method and applied it to Norwegian electricity distribution. The idea is to start with a simple DEA model that includes only a minimal set of variables that should definitely be included for theoretical or empirical reasons, and to obtain estimates for the efficiency scores with that specification. Then additional candidate variables are individually added to the model, or an already included aggregate variable is replaced with the disaggregated variables it is composed of, and in each case the corresponding efficiency score estimates are obtained. In each case a statistical test is carried out to determine whether the efficiency measures are significantly increased by the change in model specification. If so, then the candidate variable (or the proposed disaggregation) is significant, and the most significant candidate is then added to the model. This process repeats until there are no more significant candidate variables that can be added to the model.

The statistical tests are based on the assumption that inefficiency can be treated *as if* it were drawn from a random distribution. The input-oriented inefficiency for the *i*th observation is defined as: $\gamma_i = (1/\theta_i) - 1$, where θ_i is the input-oriented Farrell efficiency. It is assumed there is a density function for the inefficiencies, $f(\gamma)$, which is conditional on the output levels and the input mix. Kittelsen presents four test statistics, some of which are based on a specific assumption about the form of $f(\gamma)$ (e.g. exponential or half-normal), and one of which is a simple t-test. Besides the need for parametric assumptions, all of the tests have conceptual problems arising from the assumption that γ can be treated as an independently distributed random variable, because DEA-based scores do not have that property.[15] However, Kittelsen's simulation results suggested the tests, when used together, perform adequately in large samples. Several hundred observations, at a minimum, would be required for these tests to be reasonably reliable. However, the sample sizes used in TSO benchmarking are often smaller than that.

5.2.2 Bootstrap Tests

Simar and Wilson (2001) developed hypothesis test procedures that also test for the impact of changes in model specification on measures of inefficiencies (here the proportionate change in the distance function). But instead of assuming distributions for the inefficiencies for the purpose of formulating test statistics, they used the bootstrap method to generate an estimate of the distribution of the test statistic. Simar and Wilson's method works from more general models to more parsimonious models by testing:

- whether some outputs or inputs are irrelevant and can be safely excluded; and/or
- whether some of the inputs or outputs can be aggregated, thereby reducing the number of independent inputs and outputs.

The tests relating to narrowing the range of inputs are carried out using the output-oriented DEA model, and for tests relating to narrowing the range of outputs the input-oriented model is used. The basic test for whether some input(s) can be excluded involves first partitioning the set of inputs (x) into the 'known' (or necessary) inputs (x^1) and those that are potentially irrelevant (x^2). In this test the null hypothesis is that: $D^O(x^1, y) = D^O(x^1, x^2, y)$, where D^O refers to the output-oriented distance function. If the null hypothesis is true then output is produced only using x^1 and is not influenced by x^2. To test whether some output(s) can be safely excluded, the outputs are partitioned into the essential outputs (y^1) and the possibly irrelevant outputs (y^2) and the null hypothesis is that: $D^I(x, y^1) = D^I(x, y^1, y^2)$, where D^I refers to the input-oriented distance function. If the null hypothesis is true then the output(s)

[15] The results may also be path dependent, i.e. may depend on the order of variables being tested.

in y^2 are not relevant. The test for whether some inputs can be aggregated involves comparing two different input sets, the wider input set (x) and the aggregated input set (x^a). In this case the null hypothesis is: $D^O(x^a, y) = D^O(x, y)$.

Each of these hypotheses is tested by quantifying both sides of the equation and assessing whether differences between the estimates are large enough to cast doubt on the null hypothesis. This involves defining one or more test statistics based on the (proportionate) differences between the relative distance functions, aggregated over all firms. Bootstrapping is used to estimate the characteristics of the probability distributions, such as the critical values and p-values of these statistics. This information is used to determine the statistical significance of the computed test statistics.[16]

The Simar and Wilson method is a methodological improvement over the early Stepwise DEA formulation. However, it is computationally demanding, because the bootstrap process involves drawing a large number (e.g. 2000) pseudo samples and running the DEA model using each of those samples, for each variation in the model specification to be tested. Furthermore, these tests cannot be expected to perform well in small samples, where there is not enough data to identify the statistical influence of a candidate variable. Even in simple scenarios, Simar and Wilson found that a sample in excess of 400 would be needed if there were three variables in the DEA model (i.e. both outputs and inputs), and if the number of variables increases then the sample would need to be greater again. Recent developments in bootstrapping procedures may be more efficient, but it remains the case that methods of statistical inference within a nonparametric context require relatively large samples to be useful.

5.2.3 Discussion

These observations suggest that with the sample sizes used in past European TSO benchmarking studies, applying stepwise DEA or bootstrapping tests for variable exclusion or aggregation within the DEA framework may not be informative. In these circumstances, some or all of the following for methodologies may be feasible methods for narrowing down the candidate variables and choosing the final inputs and outputs. Although these methods are presented as alternatives, they may also be complementary:

- *Industry expertise*: In regulatory settings there is often consultation with stakeholders, including in relation to variable choice, and this process can draw on the knowledge of industry participants and experts.
- *Reliability assessments* of trial DEA results obtained when using a limited number of alternative sets of variables. The aim is to analyse the results obtained to ascertain whether any of the alternatives have implausible or unacceptable interpretations and which models produce the most stable results that are consistent with other available information. The methods of analysis may include many of the methods surveyed in chapters 6 and 7 of this report.

[16] It should be noted that there has been further progress in bootstrapping procedures since Simar and Wilson (2001). For example, Kniep *et al* (2008).

- *Other DEA-based variable selection methods*: a number of other DEA-based techniques for filtering or narrowing the set of variables to be included in a DEA model were reviewed in our report 'Selecting Cost Drivers', and at least some may be relevant or applicable in this stage of the analysis. They include:
  - Methods that rely on partial correlations between partitioned sets of candidate variables, such as the method developed by Jenkins and Anderson (2003).
  - The 'efficiency contribution measure' method of Pastor *et al* (2002), which involves comparing differently specified DEA models to determine the incremental effect of each variable on the efficiency measures of firms.
  - The regression-based approach of Ruggiero (2005), an iterative process beginning with a minimally specified DEA model, regressing the resulting efficiency score estimates on remaining candidate variables, and identifying any significant variables that might be added to the model.
- *Stochastic frontier analysis* (SFA) or other econometric model could be used as a preliminary analysis to identify the most relevant inputs and outputs. As a frontier method, SFA should *in principle* generate results for the relative efficiencies of firms that are broadly similar to the results of DEA, and therefore it may provide a useful framework for variable selection. The methods of variable selection within a regression framework are based on statistical hypothesis tests of whether certain variables have little or no effect on the dependent variable and hence can be safely excluded. There are some well-established strategies for narrowing down the variables within a regression model. The effectiveness of these methods will also be limited by small sample sizes, but arguably less so than DEA-based methods of statistical inference. Two methods that may be useful are:
  - the *general-to-specific* (GETS) modelling procedure: an algorithm designed select a parsimonious final model from a large set of variables while avoiding *ad hoc* or subjective decisions. In the GETS approach, an initial model is formulated that expresses the economic relationship being estimated in its most general form and encompasses all of the variables and effects of interest. An iterative procedure is carried out to eliminate the variables having least influence on the Bayesian Information Criterion (BIC), and using a range of other tests.[17]
  - *global search regression* is a computationally intensive algorithm which involves carrying out an exhaustive search over all possible narrower specifications, given a set of initial variables. It is designed to avoid any problem of path-dependence that can arise in iterative processes such as GETS (particularly if there is a high degree of collinearity between variables).[18] This is a powerful method, however,

[17] A procedure of this kind has been implemented in the Stata user-written program *genspec* (Clarke, 2014). The user can specify a level of statistical significance to yield a wider or narrower set of final variables. It can be employed with standard panel data models, but it may be more difficult to apply to SFA models.

[18] A procedure of this kind has been implemented in the Stata user-written program *gsreg* (Gluzmann and Panigo, 2015). It can be employed with standard panel data models, but it may be more difficult to apply to SFA

computer resources limit the number of variables that can be included, given the exponential multiplication of possible alternative specifications as the number of right-hand-side variables increases. This may not be a problem in the case of reaching a final specification from short-list of candidate variables.

- *Widening the sample*: As a general matter, it would be highly desirable to expand the sample size of TSO benchmarking studies, firstly by adding more years of data, and secondly by including a wider set of international comparators if possible. However, it may be that TSOs outside Europe are less comparable, for a range of reasons. Even so, an initial DEA modelling exercise could be conducted using a sample that included North and South American data purely for the purpose of assisting to specify the appropriate variables to be included in the model. Once the variable specification was chosen, a narrower sample (such as only the European TSOs) could then be used to estimate the regulatory benchmark efficiency measures. Using a larger sample in the variable selection process would add considerably to the ability to identify the better specifications. On the other hand, it would be more demanding in terms of data collection.
- *Principal component analysis*: Another option, discussed in more detail in our report ‘Selecting Cost Drivers’, is to use principal components analysis (PCA) to transform the set of original variables into a smaller group of derived variables that contain much of the information in the original variables, thereby reducing dimensionality with minimal loss of information, and hence minimal bias to the efficiency estimates obtained. PCA-DEA has been used in a number of DEA benchmarking studies. It involves using the leading components (or principal components) as the variables in the DEA analysis rather than the original variables. It has the particular advantages of:
    - allowing a richer set of input and output variables to be used in the overall analysis (thereby improving the ability to identify ‘true’ efficiency); while also
    - enabling a reduced number of variables used in the DEA analysis (thereby mitigating the dimensionality and discrimination problems).

5.3 Identifying Outliers

Nonparametric methods that do not have a stochastic error term, like DEA and FDH, are typically more sensitive to outliers than parametric. This can be problematic when there is some random noise in addition to the effects of the observed variables, e.g. resulting from the combined effect of unobserved factors, or to errors in the measurement of the included variables. Some outliers can determine boundary points that are not actually representative of the true production possibilities set. While there is no precise definition of ‘outliers’ (because different reasons can cause them), they can be generally defined as units that are so atypical they would not be “suitable role models in practice for setting targets for other less well-performing units” (Thanassoulis et al., 2008, p. 316). Simar has suggested that “the presence of possible outliers should lead the researcher either to identify and eliminate them … or to use stochastic frontier models if they cannot be identified” (Simar, 1996, p. 179).

models.

The super-efficiency procedures developed by Andersen and Petersen (1993), Simar (2003), Banker and Chang (2006) provide well-established methods for identifying outliers that have undue influence on the estimated efficiency frontier.[19] An outlier can only be an efficient unit (since in DEA the position of the frontier is unaffected by the presence or absence of an inefficient unit in the data sample) and each efficiency unit has a 'super-efficiency' measure, which is defined as follows. When an efficient unit is excluded from the data sample, and the DEA analysis is re-run, the frontier becomes contracted in the vicinity of the excluded unit (and other units may become efficient in that vicinity). The excluded unit is then outside the frontier and the distance from the new frontier is an indication of the 'super-efficiency' of that unit. More specifically, if the unit excluded is denoted i, and its actual level in inputs is $x_i(y_i)$ to produce outputs y_i, and after being excluded from the sample, its radial projection onto the new boundary is $x'_i(y_i)$, then the super-efficiency measure is: $x'_i(y_i)/x_i(y_i) > 1$.

Generally, efficient units will have super-efficiencies greater than one, but will vary in size. Outlier identification usually involves selecting among those units that have the highest super-efficiencies. The criteria used to identify these most extreme points may be fairly arbitrary or subjective as typically most of outlier detection methods are. For example, it may be a specific cut-off value or it may be a percentage of those with the highest values above some threshold. "The rationale is that the boundary is drawn at a level of performance that a significant percentage of units can attain and so that is deemed an acceptable benchmark in practice" (Thanassoulis et al., 2008, p. 318).

A second approach is to examine the effect of including and excluding a firm from the sample on the average efficiencies of the remaining units. A problem with this approach is that it measures the influence of the firm, rather than whether it is an outlier, and influence is not in itself a problem. For example, a firm can have high influence if it is close to a large number of DMUs, and dominates them, whereas an outlier is an atypical firm which is far from the other units (Thanassoulis et al., 2008).

Germany's electricity distribution network regulation has two criteria for identifying outliers, one based on the requirement that the DSO should not have too large an influence on the average efficiencies of other DSOs, and the other being that DSOs that are extremely superefficient should be excluded (i.e. if above a particular multiple of the interquartile range of super-efficiencies). Outliers identified by either of these criteria should be excluded (Agrell and Bogetoft, 2016b).

Once outliers are determined, there is the question about how to deal with them. Although some authorities recommend automatically eliminating them from the sample, it is advisable to firstly better understand what they represent. Thanassoulis et al emphasise both the need to be careful about how outliers are dealt with, and the need for transparency:

> … the detection of outliers and deciding what to do with them are two separate issues. One should always check the data for outliers since these can indicate some errors in data measurement or simply some observations that are atypical (and this should be reported in

[19] Outliers may also be observations that appear to be exceptionally inefficient. These can cause problems in second stage regression analysis.

> the analysis). If the outliers are defining the shape of the efficiency frontier and influencing by a large degree the efficiency of the other units, then one can consider an assessment with and without the outliers and compare the results. " (2008, p. 319).

Recently, Simar and Zelenyuk (2011) developed a method to take care of outliers and general noise in a systematic, rather than an *ad hoc*, way. It involves two stages: first filtering away outliers and other noise identified by estimating a nonparametric stochastic frontier model, and then using DEA (or FDH) on the filtered data. They referred to it as Stochastic DEA (or Stochastic FDH). Rather than removing outlier observations, this method adjusts them by removing the disturbance component. Since the method involves nonparametric SFA at the first stage, it generally demands relatively large samples, especially if the number of inputs and outputs is large.

5.4 Regulatory Considerations

Objectives within the regulatory framework will also be relevant considerations in the selection of a preferred model. This is because the benchmarking model, and the targets generated by it, may have an influence on the incentives of regulated businesses. For example, it may be desirable that the outputs reflect a balance of regulatory priorities so that businesses have incentives to achieve targets most relevant to the regulatory framework. Agrell and Bogetoft have emphasised that more research is needed into how to identify and quantify the incentives for regulated businesses produced by different benchmarking models, which would assist to compare the likely effectiveness of alternative models if applied (Agrell and Bogetoft, 2016b, p. 33).

According to Agrell and Bogetoft (2016b), the choice of a preferred benchmarking model (or models) involves the application of multiple criteria. These include not only objective considerations such as whether the model performs relatively well from a purely technical or statistical perspective and whether it conforms to conceptions drawn from economic theory or industry knowledge, but also involves some broader considerations. These may include the subjective assessment of expert practitioners and issues of regulatory process or application.

These criteria are not well defined and these authors emphasise that there is:

> … a need for more serious academic discussions of what a good regulatory benchmarking model actually is … the actual choice of the model invokes a series of conceptual, statistical and pragmatic criteria, and it is not clear how to prioritize and weight different concerns. (Agrell and Bogetoft, 2016b, p. 33)

6 TESTING THE REPRESENTATIVENESS OF A MODEL

A common issue that arises in regulatory settings is whether the results of efficiency analysis are sufficiently representative for a particular business that they can be relied on to draw inferences about the scope for efficiency improvement in that business. The regulator has an interest in the reliability of the results to ensure that regulatory settings are consistent with both the efficiency and viability of the regulated businesses. The regulated businesses will generally have incentives to claim there are aspects about their business which are sufficiently different to other businesses in the sample to cast doubt on the reliability of the results, and so thereby to obtain greater 'information rents' due to greater uncertainty around their efficiency potential. This section discusses some methods that can be used to assess the degree of reliability of the results of DEA analysis and discusses a number of methods for adjusting efficiency estimates to make them more representative. These include:

- approaches to comparing the results to other sources of information and to other models estimated using different methods are discussed in section 61.
- information on the robustness of the DEA efficiency estimates is also an important part of their proper interpretation. Methods of assessing robustness are discussed in section 6.2.
- adjustments to take account of 'slacks', which are sources of inefficiency not taken into account in the conventional radial efficiency estimates (section 6.3), and
- adjustments for potential sampling bias, due to the risk that the set of firms included in the sample may not include some efficient firm that, were it included, would influence the estimate of 'best practice' and hence alter the efficiency scores of other firms (section 6.4)
- identifying and removing the effects on measured efficiencies of exogenous operating environment factors that affect the firms in the sample differently, and are not related to the performance of the firms and cannot be influenced by actions management can take (section 6.5).[20]

6.1 Comparing results to other sources

6.1.1 Comparing results to previous studies

A general question in evaluating the results of a model is whether the efficiency scores and rankings obtained from the analysis are consistent with other available information, which can include previous benchmarking studies or the views of experts with more detailed knowledge of the operating practices of the businesses being compared. When results are

[20] By 'cannot be influenced by' we mean that the operating environment factors are exogenous for the firm. Management can still make choices in how to deal with operating environment factors (which may be more or less effective), but these responses generally require resources to implement, so that differences in operating environments can affect the observed comparative productivity and cost efficiency of firms even when action is taken to mitigate their effects. The effect of operating environment factors is an empirical question.

inconsistent with other sources of information, then further analysis is warranted to understand the results in more detail so that the benchmarking model can be critically appraised. This is essentially the subject of the following sections of this chapter.

6.1.2 Comparing results to parametric and index methods

Greene has suggested that it "is always interesting to compare the DEA results with those obtained using the stochastic frontier model" (Greene, 2007, pp. E33-118). The technical efficiency scores obtained using the two methods can then be plotted and the degree of correlation assessed. If a TSO of interest is an outlier on a plot of this kind, this may indicate that the DEA score for that business may be comparatively unreliable. Considerations of this kind will be relevant not only to the measured efficiency of the TSO being regulated, but also for the group of efficient TSOs against which it is being compared. Comparison of efficiency rankings obtained using DEA and SFA may also be instructive. It is also possible to compare the results of DEA modelling with other techniques such as multilateral total factor productivity indexes (MTFP).

6.2 Sensitivity analysis

An approach often used is to test the variability of efficiency measurements, or whether similar results are obtained for rankings when using different DEA models—e.g. different methodologies, or different ways of transforming the variables, or under variations in the data. Bootstrapping methods can be used to examine the confidence intervals for measures of efficiency. This section discusses methods used to analyse the sensitivity or stability of DEA efficiency results due to variations in the data.

Seiford and Zhu (1998) review several approaches to sensitivity analysis that involve perturbing the data to establish just how much data error is needed to substantially change an efficiency score (either rendering an efficient DMU inefficient or vice versa). In the approach they advocate, the data is perturbed for all DMUs across different subsets of inputs and outputs. The emphasis here is on the effect of changes in the data for inputs and outputs on the binary result concerning whether a DMU is, or is not, efficient. Stability relates to the magnitude of the changes to the data that are needed to disturb that basic result. This differs from the bootstrap approach, discussed in section 3.2, which quantifies the sensitivity of efficiency scores to changes in the DMUs included in the sample (rather than changes in the data) for the purpose of constructing confidence intervals.

Cooper et al (2004) survey several methods of sensitivity analysis. Only one is discussed briefly here, which they describe as the 'envelopment approach' because it is based on the DEA envelopment model. The aim is to "identify allowable variations in every input and output for every DMU before a change is status occurs for the DMU_o being analyzed" (Cooper et al., 2004, p. 89). These methods use linear programs that are variations on the basic input- and output-oriented DEA models, designed to solve for the boundary values of each input and output that are just sufficient to change the status of DMU_o from efficient to inefficient or vice versa (the program is solved once for each variable for each DMU). This "work has moved from evaluating one input or one output at a time in one DMU and has proceeded into more general situations where all inputs and outputs for all DMUs can be simultaneously varied" (Cooper et al., 2004, p. 95). The envelopment approach has been implemented in accessible Excel-based software (Zhu, 2003).

Sensitivity analysis of this kind often tends to show that quite substantial changes in inputs or outputs would be needed to change the status of a DMU—that is, the DEA result is robust. In any event, it would be advisable to have information that sheds light on how robust the DEA findings are.

6.3 Taking account of slacks

When making efficiency comparisons it is appropriate to take into account any 'slacks' (when a business lies on a vertical or horizontal segment of a frontier). Whenever there are multiple outputs (in an output-oriented setting) or multiple inputs (in an input-oriented setting), then it can be that some DMUs that are on the efficiency frontier are not actually fully efficient because of the presence of 'slack'. Fried et al observe that a "notable feature of the Debreu-Farrell measures of technical efficiency is that they do not coincide with Koopman's definition of technical efficiency" (2008, p. 25). They "only require the absence of radial improvements" but can be strictly dominated by a point on the efficient section of the input isoquant due to slackness in the use of at least one input. It is therefore important to examine slacks in addition to efficiency scores.

Some of the efficient firms at the solution of a conventional DEA program may have one or more 'slacks' because the reduction of inputs (in the input-oriented case) or the expansion of outputs (in the output oriented case) is restricted to radial (i.e. equiproportionate) reductions or expansions as the case may be. For example, if the input-oriented model is used, the DEA (Farrell) efficiency score represents the proportion by which all inputs could be equiproportionately reduced (i.e. preserving the mix) while still allowing an efficient firm to produce the same output vector with the same technology. Pareto (or Koopmans) efficiency is a higher efficiency standard than Farrell efficiency. A given input vector *x* is Pareto efficient for producing a given output vector *y* if, with a reduction in *any element* of *x*, it would no longer be feasible to produce *y* with the same technology. When input reductions are confined to radial contractions there still may be one or more inputs that could be reduced while producing the same set of outputs. Hence, some firms that are Farrell efficient may not be Pareto efficient (but all Pareto efficient firms are Farrell efficient also). In other words, the set of firms that are Pareto efficient is, in general, a subset of the firms that are Farrell efficient. Pareto efficiency is the more useful measure for the purposes of economic regulation.

Therefore, attention needs to be given to slacks. The presence of any positive input or output slacks at the optimal solution of a DEA model means there is a potential problem with the technical efficiency (TE) measures. A firm that has a TE = 1, but has slacks in some input(s) is not fully efficient, and comparisons of TE between two firms will not be valid if one has slacks. Some studies report slacks alongside TE scores but make no adjustment to the TE score. Perhaps a better practice is to report, in addition to the original TE estimates and the slacks, an adjusted TE measure which takes account of the slacks.

There are various methods of adjusting the TE score for the removal of slacks. This topic was introduced in section 3.4 in the context of identifying Pareto efficient firms. In one of the methods for removing slacks outlined in section 3.4, initial estimates of TE scores are obtained using the conventional input-oriented DEA model. The radial targets (or projection points on the efficiency frontier) implied by those TE scores (ie each firm's original input vector multiplied by its efficiency score) are used in place of the actual inputs for inefficient firms in a second-stage DEA analysis, this time using the non-radial Russell efficiency method. This approach yields a modified efficiency measure, which is the product of the Farrell efficiency measure from the first stage, and the Russell efficiency measure from the second stage.

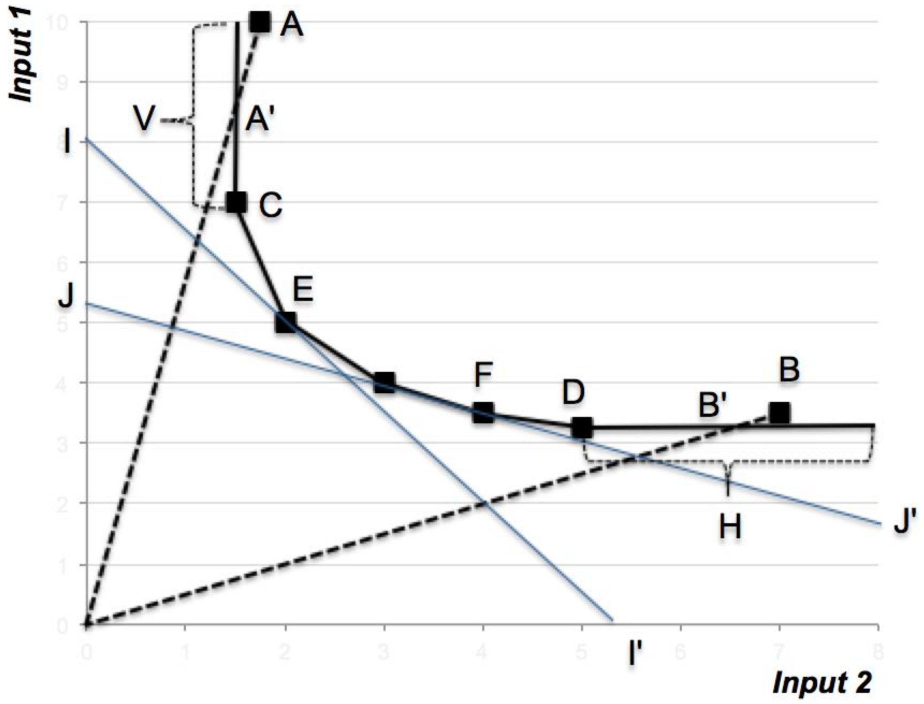
Ray (2004) presents another approach which uses the assurance region (AR) method of imposing constraints on DEA multipliers. This approach is based on the fact that when a slack is present at the optimal solution, the relevant input or output constraint is nonbinding, and the multiplier (or shadow price) of that input or output equals zero. By imposing constraints on ratios of the input multipliers and on ratios of the output multipliers, then this (and the other constraints in the program) will ensure that none of the weights can take zero values, and hence the solution cannot include slacks. This approach involves a single stage in principle (although in practice some experimentation may be needed to determine the multiplier constraints).

Both of these methods produce a revised measure of technical efficiency that should be more suitable for benchmarking purposes. As previously stated, measures of technical efficiency that incorporate slacks can provide misleading comparisons. However, it should be noted that measures of cost efficiency cannot include slacks, because in the cost efficiency program it is assumed that all inputs have a positive price, and therefore any cost-minimising combination of inputs must be a Pareto Efficient combination (assuming no output allocative inefficiency or output slacks). If comparisons are made on the basis of cost efficiency alone, then the issue of slacks does not arise. This is shown in Figure 6.1.

The DEA-estimated input isoquant is shown as a solid dark segmented line. The extreme segments of this line are labelled as V and H. These segments reflect the free disposability assumption, so V is vertical and H is horizontal. Interior points that are projected onto either of these two segments have a slack. Inefficient input mix A is projected onto point A' on the frontier, with a Farrell efficiency score equal to 0A'/0A and a slack equal to the line segment A' to C. Similarly, inefficient input mix B is projected onto point B' on the frontier, with a Farrell efficiency score equal to 0B'/0B and a slack equal to the line segment B' to D. The segments of the frontier between C and D are all Pareto technically efficient points.

While there are many technically efficient points, for a given set of input prices there is only one input mix that minimises cost in this example (although, in general, there are could be many optimal input mixes). Two alternative sets of input price lines are depicted in Figure 6.1. Firstly, with input price line I to I', point E represents the cost minimising set of inputs. Secondly, if line J to J' is used, then F is the cost minimising input mix. Because input prices are assumed to be positive, the input price line cannot be vertical or horizontal and therefore can only be at a tangent to cost-efficient points.

Figure 6.1: **Input Isoquant**



Source: Economic Insights.

6.4 Adjusting for Bias

DEA estimates of technical efficiency are known to be *consistent*, that is, for increasingly large data samples they ultimately converge toward the ‘true’ efficiency levels. However, the rate of convergence may be slow for large dimensions, so data samples need to be quite large before there can be a high degree of confidence in the accuracy of the estimates. A second issue is bias in the estimated efficiencies. DEA efficiency estimates always have some bias in finite samples, although this bias becomes very small in large samples. This bias is because the frontier is estimated from the most efficient firms in the data sample, and is the tightest fitting convex linear surface that envelops the data of those firms. As the dataset expands to include more firms, there is always a small (perhaps very small) positive probability that one of the firms added to the sample will be more efficient, and render inefficient one of the previously efficient firms. This amounts to a positive finite probability that the frontier will be widened as the sample increases, and therefore efficiency scores estimated with finite samples will, in probability, be over-estimated (for input-oriented efficiency scores). This statistical theory is based on the assumption that any given dataset is a randomly drawn sample from a population.

The bootstrap method can be used to estimate the bias of the efficiency scores. Bootstrapping involves drawing a large number of random ‘pseudo samples’ of observations from the existing dataset (with replacement). If the existing dataset has *n* firms, then each pseudo sample usually has *n* or less observations, and a very large number of such samples can be drawn. The theory underlying the bootstrap is that this form of resampling from the dataset is a valid proxy for resampling from the population.[21]

The estimated bias of an efficiency estimate, obtained from the bootstrapping procedure is itself subject to statistical error, and the variance of that error may be large, relative to the estimated bias, in small samples. Simar and Wilson state that the variance of the bias estimate is "typically of smaller magnitude" than the bias estimate "for reasonable sample sizes" (2007, p. 39). That is, for reasonably large samples. In the context of two-stage DEA analysis, Simar and Wilson provided evidence, using simulation experiments, that for various scenarios with a simple model with a total of three variables (e.g. 2 inputs and 1 output) bias correction only became worthwhile (for the purpose of using bias corrected scores in second-stage analysis) with a sample size of 800. When more variables are used in the DEA model, a larger sample size *may* be needed for bias correction to be worthwhile before undertaking subsequent analysis using the efficiency estimates (e.g. second-stage DEA).[22]

These observations suggest that, for the sample sizes commonly used in utility benchmarking studies, bias adjustment of efficiency scores is not likely to be accurate enough to be warranted until either larger samples are collected or bootstrap methods are improved.

6.5 Adjusting for Operating Environment Differences

As discussed in section 3.8, there will often be 'operating environment' or 'contextual' variables that are outside the control of the DMU but influence the ability of a firm to translate inputs into outputs given the available technology. It is desirable if not essential to make adjustments for such influences.

> Taking the heterogeneity of firms and their operating environments into account is important in virtually all thinkable empirical applications of productive efficiency analysis. If the heterogeneity is ignored, firms operating under favourable conditions appear more efficient than firms operating in a harsh environment. (Johnson and Kuosmanen, 2012, p. 560)

Section 3.8 also discussed alternative methods of adjusting for exogenous factors that influence the available production possibilities. The most popular method is the 'two-stage' approach in which DEA efficiency scores are estimated in the first stage and in the second stage those efficiency scores are regressed against a number of variables that measure aspects of the operating environments of the businesses. This section discusses methodologies for second-stage regression analysis of DEA efficiency scores.

The aim of the second-stage analysis is to establish how much of the observed differences in efficiency can be explained by operating environment factors. The parameters of the

[21] In the DEA context, Simar and Wilson (1998) showed that when the pseudo-sample is of the same size as the original sample and drawn via the empirical distribution function, the estimates are not consistent. They proposed a practical bootstrap procedure based on kernel smoothing of the distribution of efficiency, and Kneip *et al* (2008) improved on the sub-sampling bootstrap for DEA and FDH and proved that it provides consistent estimates under fairly general conditions.

[22] The required sample size is a complex matter that depends on the scenarios being tested.

estimated second-stage model can be used to adjust for the effects of the operating environment variables from the estimated efficiency scores, to provide a more reliable basis for measuring differences in efficiency due to business decisions and performance. Box 6.1 explains the nature of the adjustment to the efficiency scores based on the results of the second-stage regression.

Box 6.1 Controlling for Operating Environment Characteristics

Suppose the 2nd stage regression is:

(1) $$\hat{\theta}_i = \alpha_0 + \sum_h \alpha_h z_{ih} + \varepsilon_i$$

where $\hat{\theta}_i$ is the estimated cost efficiency for firm i from the 1st stage (DEA) analysis, z_{ih} is the value of operating environment characteristic h for firm i, and ε_i is random noise. Let $\tilde{\alpha}_0$ and $\tilde{\alpha}_h$ (h = 1 … *H*) be the estimators obtained from carrying out regression (1). Then the corrected efficiency score estimate, θ_i^*, can be obtained from:

(2) $$\theta_i^* = \hat{\theta}_i - \sum_h \tilde{\alpha}_h (z_{ih} - \bar{z}_h)$$

where $\bar{z}_h$ is the sample mean value of operating environment characteristic h.

The benefits of the second stage analysis are not limited to adjusting efficiency scores. This analysis also helps to explain the results of the first stage analysis.

> Estimating the effects of contextual variables on efficiency can provide valuable insight to managers who develop business strategies or make decisions on operating practices, and for policy makers who may influence the external operating environment of firms through standards, regulations, taxes, subsidies, and other policy measures. (Johnson and Kuosmanen, 2012, p. 559)

Perhaps the most popular approach to the second-stage regression has been Tobit-regression, which is a particular case of censored regression. That is, it allows for some bunching of observations of the dependent variable at a boundary value, and in the case of second-stage regression the efficiency scores are bounded at one from above (or from below in some formulations of the output-oriented efficiency score).[23] This approach has been subject to criticism, as discussed below. Others, such as McDonald (2009), have viewed the efficiency scores as fractional data and used a fractional outcome regression method, such as logit, probit or more general methods. A possible difficulty is that fractional outcome regression is often used with data in which the extreme observations (0 and 1) are rare, which is not the case with efficiency scores.

Simar and Wilson (2007) argued that the censored regression model was inappropriate

[23] In the input-oriented model, efficiency scores are also, in principle, bounded from below at zero, but any bunching of efficiency scores at zero is highly unlikely, so this can be safely ignored.

because, although there is usually some bunching of estimated efficiency scores at the boundary where the efficiency scores equal one, this is only an artefact of small samples. For reasons explained in section 6.4, the efficient firms in the data sample may actually fall short of the ‘true’ frontier, if the sample were infinitely large. They proposed using truncated regression after removing all of the firms found to be efficient, based on their finding that without accounting for truncation the second-stage estimator is biased and inconsistent. Unlike the Tobit model, in which all of the efficiency scores are used in the second-stage analysis, the Simar and Wilson method requires efficient scores be removed from the data set before undertaking the second-stage analysis. The resulting loss of degrees of freedom may be a limitation of this approach in the case of small samples.[24]

Simar and Wilson (2007) developed a single bootstrap procedure for regressing DEA efficiency scores against the exogenous factors, and a double bootstrap procedure, which regresses the bias-corrected efficiency scores against the exogenous variables. Both of these procedures take account of the complex serial correlation between efficiency scores. As discussed in section 2.1, unless sample sizes are quite large, it may not be advantageous to use the bias corrected efficiency scores in the second-stage analysis. In that case, the single bootstrap procedure can be used, which addresses only the problem of serial correlation in the efficiency scores.[25]

The single bootstrap procedure involves, firstly, estimating the DEA model for each firm to obtain the set of efficiency score estimates. These estimates are then used as dependent variables in the second-stage (maximum likelihood) regression. Then a large number of values are randomly drawn from a truncated normal distribution, where the truncation point and the variance of the distribution are estimated from the initial parameter estimates. These are used to construct more appropriate confidence intervals for each of the parameters previously estimated.

More recently, Johnson and Kuosmanen (2012) re-examined the issue of consistency and bias of the second-stage regression estimator, and found that “the OLS regression model of the DEA efficiency scores on the contextual variables provides a statistically consistent estimator of the coefficients of the contextual variables” under reasonably general assumptions (2012, p. 560). However, if there is bias in the efficiency score estimates it will carry over into the second-stage regression, although bias in DEA scores tends to disappear in larger samples. Although their findings seemed to suggest that the second-stage regression method is adequate, these authors developed and preferred a single-stage approach in which the operating environment variables are incorporated into the first-stage model of the efficiency frontier. However, their proposed one-stage method is an application of nonparametric least squares convex regression, which is a substantial departure from conventional DEA analysis.

Within the context of DEA analysis, the two-stage approach to addressing operating

[24] Fewer observations may be removed in the double bootstrap procedure discussed in the following paragraph because bias correction renders previously efficient businesses inefficient.

[25] This method has been implemented in a user-written Stata program: ‘simarwilson.ado’. Harald Tauchmann, ‘simarwilson: DEA based Two-Step Efficiency Analysis’ (June 26, 2015: German Stata Users Group Meeting).

environment variables has considerable merit, and the second-stage regression methods proposed by Simar and Wilson appear to be sound. Although Simar and Wilson only considered linear relationships between the efficiency scores and the operating environment variables, the method may be applicable to other functional forms (such as long-linear or log-log functional forms). One issue not discussed by Simar and Wilson is the appropriate treatment of slacks in the second stage regression. We would expect that the efficiency scores should be adjusted to take slacks into account before using them in the second-stage regression analysis, although this is not yet resolved in the DEA literature.

7 FURTHER ANALYSIS AND INTERPRETATION

The previous chapter discussed methods of testing the representativeness of the results of a model and adjustments to efficiency measures that may be needed to make them more useful for benchmarking purposes. The results of a DEA study can also be used to derive a range of other quantitative information, which can assist to interpret the efficiency findings, which is the subject of this chapter.

Types of information that can be obtained through further analysis include:

- estimates of the relative importance of different types of inefficiency, such as technical and allocative efficiency (section 7.1);
- changes in inefficiency over time, and the relative importance in the observed productivity changes of factors such as 'frontier shift' and 'catch-up' to the efficiency frontier (section 7.2 and 7.3);
- elasticities of substitution between inputs and other quantitative data that describes the economic characteristics of the estimated production possibilities set (Section 7.4).

7.1 Decomposing cost efficiency

The efficiency score obtained from the DEA cost minimisation model measures the degree to which costs could be reduced while still producing the same outputs. Cost efficiency (*CE*) is defined as the ratio of the optimum (minimal) cost to the actual cost. If the input-oriented DEA technical efficiency model is also computed, then cost efficiency can be decomposed into two parts, technical efficiency and allocative efficiency.[26]

Input-oriented technical efficiency (*TE*) measures the degree to which inputs can be radially contracted (that is, preserving the mix) whilst still producing the same outputs. Allocative efficiency (*AE*) measures the cost efficiency improvement that could be made solely by changing the mix of inputs. The DEA cost minimisation program produces a measure of *CE* for each firm, and the input-oriented DEA technical efficiency program produces a *TE* measure for each firm. The *AE* for each firm can then be calculated using the definition: $CE = TE \times AE$.

This decomposition helps to explain the sources of inefficiency and is important information for the management of firms, because strategies directed to reducing technical inefficiency may differ from the strategies needed to reduce allocative inefficiency.

7.2 Time-period Comparisons: The Malmquist Productivity Index

Where there are several years of data for the DMUs in the sample, the DEA analysis can be conducted using different periods to assess the stability of the results, and the plausibility of changes in measured efficiency over time. Assessing the stability of model results over time is particularly important when small samples of firms are used in the benchmarking analysis.

[26] This assumes there is more than one input.

When panel data is used in DEA analysis,[27] information can be obtained on changes in total factor productivity (TFP) for each DMU between each period included in the sample. The average productivity change for the sector as a whole can also be calculated. This section briefly describes how productivity changes are calculated from DEA results.

The Malmquist productivity index (MPI) is a measure of total factor productivity (TFP) change based on efficiency scores of the kind calculated in DEA programs. The MPI for DMU_o is the product of two components: (i) a measure of its 'catch-up' to the efficiency frontier; and (ii) and measure of 'frontier shift'. Despite it being formulated by reference to relative movements in the efficiency frontier, and in DMU_o's input-output mix, it has been shown that under certain assumptions about the production technology, it is equivalent to a measure of TFP based on input and output indexes calculated using the Törnqvist index number formula, widely used by statisticians for price and quantity indexes (Caves et al., 1982b).

In conventional DEA analysis, the efficiency score for each firm is solved using a single period. When there is more than one period in the dataset, this procedure can be carried out separately for each year in the sample, thereby estimating an efficiency score for each unit in each year. For DMU_o, the procedure just described produces, for each period, an efficiency measure calculated using its input-output mix in that period assessed against the efficiency frontier for the *same period*. For example, let (x_o^s, y_o^s) be its input-output mix in period *s*, which represents the coordinates of a point, which is compared to the efficiency frontier in the same period to obtain the efficiency score: $\theta^s(x_o^s, y_o^s)$. Here the superscript applied to theta refers to the date of the technology (or frontier) against which the input-output mix is compared. In principle, the input-output mix in period *s* could also be compared to an efficiency frontier in another period, say *t*. In that case we would refer to it as $\theta^t(x_o^s, y_o^s)$.

This notion of comparing the input-output mix of one period to the technology of another period is central to the calculation of the MPI. This is because the MPI seeks to measure the change in DMU_o's efficiency score resulting from the change in its input-output mix *when calculated against a fixed technology*.[28] There are two natural ways to measure this, and the MPI is the geometric average of these two measures:

(3.1) $$MPI = \sqrt{\Delta^s . \Delta^t} \quad \text{, where:}$$

(3.2) $$\Delta^s = \frac{\theta^s(x_o^t, y_o^t)}{\theta^s(x_o^s, y_o^s)} = \frac{\text{Efficiency of } (x_o^t, y_o^t) \text{ assessed against period } s \text{ frontier}}{\text{Efficiency of } (x_o^s, y_o^s) \text{ assessed against period } s \text{ frontier}}$$

(3.3) $$\Delta^t = \frac{\theta^t(x_o^t, y_o^t)}{\theta^t(x_o^s, y_o^s)} = \frac{\text{Efficiency of } (x_o^t, y_o^t) \text{ assessed against period } t \text{ frontier}}{\text{Efficiency of } (x_o^s, y_o^s) \text{ assessed against period } t \text{ frontier}}$$

[27] When a dataset used in analysis spans more than one period whilst covering the same DMUs, it is referred to as panel data.

[28] The MPI is more commonly defined as the ratio of two distance functions. However, distance functions are inversely related to DEA efficiency scores. The presentation here avoids the need to introduce the concept of distance functions.

The denominator of (3.2) and the numerator of (3.3) are available from the DEA program computed for each individual year of the sample period. However, the numerator of (3.2) and the denominator of (3.3) are intertemporal calculations that are not available from that program. A related DEA program needs to be executed to compute the intertemporal efficiency scores (see: Tone, 2004). The within-period and intertemporal efficiency scores are then used to compute the MPI.

Estimates of changes in TFP obtained by calculating of the Malmquist productivity index are valuable information for several reasons:

- First, information on productivity trends can be a useful diagnostic check on the benchmarking model. The results for the Malmquist productivity index can be sensitive to the choice of variables included in the model. If there is other information relevant to productivity trends in the industry, comparison with the model-implied productivity trends can be a useful test of the plausibility of the model.
- Second, the performance of DMUs can be compared to their own past performance (and not only compared to the efficiency frontier). Estimates of TFP change for an inefficient DMU can shed light on whether they are making progress in moving toward the efficiency frontier (or attaining efficiency), or on the other hand whether they are regressing and becoming inefficient, or more inefficient.
- Third, the changes in productivity of one DMU can be compared to the change in productivity of other DMUs in the sample, thus providing another dimension to benchmarking. This information can show how the productivity change of a DMU compares to those of other DMUs, particularly those it is most closely comparable to. It can also shed light on whether there have been any substantial changes in the groups of DMUs found to be on the frontier from one period to another.
- Fourth, there are several useful ways in which changes in the Malmquist productivity index can be decomposed into separate explanatory factors, including technical change, changes in technical efficiency, and the effects associated with changes in output levels when there are variable returns-to-scale. These methods are discussed in the following section (7.3). More generally, information on productivity changes, and their decomposition, may shed light on sustained patterns or trends that may assist to interpret measured efficiency scores.

In regulatory applications, the calculation of productivity trends has a further usefulness. If the regulatory cost targets embodied in the revenue cap are to incorporate an allowance for industry-wide productivity gains associated with frontier shift, that estimate would need to be obtained from some source. There is greater internal consistency to the methodology if it obtained from the same dataset used to quantify firm-specific inefficiencies. Furthermore the MPI decomposition (discussed in the next section) allows the frontier shift effect which is common to all firms in the sample, to be separated from other sources of productivity change, such as firm-specific changes in inefficiency. To reiterate, the computation of Malmquist TFP indexes requires a panel dataset.

7.3 Decomposition of MPI

The MPI not only provides information about changes in TFP between periods. It can also be decomposed into several sources of productivity change to identify the relative importance of the main contributors to that change, including:

- changes in a DMU's degree of inefficiency, or 'catch-up'
- technical change or 'frontier shift', and
- the effects of changes in scale efficiency resulting from changes in demand.

This information can be useful in gaining a greater understanding of the nature of changes in the productivity of a DMU, which assists to explain its prevailing efficiency or inefficiency relative to other firms in the benchmarking sample. This section briefly explains the main elements of the decomposition of the MPI (following Tone, 2004). The MPI is equal to the product of 'catch-up' and 'frontier shift' effects defined as follows.

(3.4) $$MPI = '\text{Catch}-\text{up}' \times '\text{Frontier shift}'$$, where:

(3.5) $$'\text{Catch}-\text{up}' = \frac{\text{Efficiency of } (x_o^t, y_o^t) \text{ assessed against period } t \text{ frontier}}{\text{Efficiency of } (x_o^s, y_o^s) \text{ assessed against period } s \text{ frontier}}$$

(3.6) $$'\text{Frontier shift}' = \sqrt{\varphi_s . \varphi_t}$$, where:

$$\varphi_s = \frac{\text{Efficiency of } (x_o^s, y_o^s) \text{ assessed against period } s \text{ frontier}}{\text{Efficiency of } (x_o^s, y_o^s) \text{ assessed against period } t \text{ frontier}}$$

$$\varphi_t = \frac{\text{Efficiency of } (x_o^t, y_o^t) \text{ assessed against period } s \text{ frontier}}{\text{Efficiency of } (x_o^t, y_o^t) \text{ assessed against period } t \text{ frontier}}$$

The catch-up term is simply the ratio of the within-period efficiency scores from periods *t* and *s*. The frontier shift term is the geometric average of two measures of the impact that the change in the frontier from period *s* to period *t* has on the measured efficiency of a fixed input-output mix. DMUo's input-output mixes in periods *s* and *t* are each used as fixed reference points.

The forgoing decomposition is based on the DEA variable returns to scale (VRS) program, with its associated Malmquist productivity index MPI_V. When the constant returns to scale (CRS) program is used instead, the associated productivity index, MPI_C has a slightly different decomposition which reveals information about effects related to changes in scale efficiency. The two MPI measures are related as follows:

(3.7) $$MPI_C = '\text{Catch}-\text{up}' \times '\text{Frontier shift}' \times \text{ Scale efficiency change}$$

(3.8) $$= MPI_V \times \text{ Scale efficiency change}$$

Scale efficiency is measured as the ratio of the efficiency score for a particular input-output mix when CRS is assumed to the efficiency score for the same input-output mix when VRS is assumed. The scale efficiency change is again measured by a geometric mean of two measures of the change in scale efficiency, measured from frontiers for periods *s* and *t*.

7.4 Analysis of multipliers

The implicit weights of a DEA model reflect marginal rates of transformation between inputs and outputs, and if there are multiple inputs, the marginal rates of substitution between them. Benchmarking studies of energy networks carried out to date provide a body of international evidence on their cost structures, including in some case, information about marginal rates of transformation and substitution. One approach to assessing the plausibility and reliability of a benchmarking model may be to interpret the implicit weights obtained for the variables in the multiplier DEA analysis, and compare them to previous findings in the literature and expert opinions. This section discusses elasticities of substitution between inputs.

The DEA technical efficiency model estimates the PPS as a piecewise linear surface, and the slopes of this surface in particular directions reflects the opportunities to trade off the use of one type input against the use of another (keeping all other inputs unchanged). A useful method of calculating the elasticities of substitution between inputs is presented in Schmitz and Tauchmann (2012), and this method is briefly summarised. The multiplier form of the DEA program is used (here the input-oriented version).

When applied to a particular firm (DMU$_0$), the multiplier program finds a set of optimal output and input weights or shadow prices (u_l for each output *l*, and v_m for each input *m*) that:

- maximise $\sum_{l=1}^{L} u_l y_{l0}$ (total output) while $\sum_{m=1}^{M} v_m x_{m0} = 1$ (total input);
- satisfy the following constraint for each firm *i* (including DMU$_0$): $\sum_{m=1}^{M} v_m x_{mi} - \sum_{l=1}^{L} u_l y_{li} \geq 0$; and
- all *u*'s and *v*'s are non-negative.

The solutions are $(u_1^*, \dots, u_L^*, v_1^*, \dots, v_M^*)$ and at the solution there is at least one binding constraint: $\sum_{m=1}^{M} v_m^* x_{mk} - \sum_{l=1}^{L} u_l^* y_{lk} = 0$. *In principle,* an estimate of the marginal rate of substitution between inputs m and q (*MRS$_{m,q}$*) can be obtained by implicitly differentiating this equation to obtain: $MRS_{m,q} = -\partial x_{mk}/\partial x_{qk} = -v_q^*/v_m^*$. However, to obtain reliable estimates of the MRS it is necessary to take slacks into account.

The method used by Schmitz and Tauchmann (2012) is to find, for each inefficient DMU, the Pareto Efficient point on the frontier that it is projected onto (that is, the radial projection point adjusted for the slacks). A method of identifying these points was discussed in section 3.4. Once the coordinates of these efficient points are obtained, they are substituted for the actual data for the inefficient units in the sample, and the DEA problem is re-estimated. This yields more reliable estimates of the *u*'s and *v*'s for the purpose of calculating the elasticities of substitution.

Schmitz and Tauchmann (2012) also observe that the concept of elasticities of substitution needs to be adjusted to accommodate the piecewise linear production frontier, which does not have smooth curvature, so they use the concept of *technical elasticity of substitution*, which needs to be estimated using the Pareto Efficient projected points previously mentioned:

$$\widehat{TES}_{m,q} = -\frac{v_q^*}{v_m^*}\frac{x_q^{opt}}{x_m^{opt}}$$

The authors note that this technical elasticity of substitution concept is purely dictated by the technology and doesn’t rely on assumed optimising behaviour, such as cost minimisation. They used the median values of the estimates of the TES.

Quantitative information on the properties of the PPS, such as the technical elasticities of substitution between inputs, is useful information to present to users of the benchmarking analysis to assist in the understanding and interpretation of the results. This information can be compared to expert opinions on the characteristics of the technology, and to previous findings in the literature on the cost structure and marginal rates of transformation and substitution of energy networks.

8 COMBINING MODEL RESULTS

As noted by Farsi et al, an important problem faced by regulators that conduct benchmarking "is the choice among several or legitimate benchmarking models that usually produce different results" (2007, p. 1). This section addresses the topic of whether only one 'preferred' model should be relied on or whether the results of several well-performing models should be combined somehow. It also discusses some ways of combining results, including alternative methods that have been used or suggested in the literature.

Some regulators formally combine more than one approach. For example, the Energy Market Authority of Finland has at one time used an average efficiency measure obtained using DEA and SFA methods, and the German regulator has used the maximum of the efficiency measures obtained using these methods (subject to a minimum value of 0.6) (Kuosmanen, 2012). Some researchers have proposed more complex algorithms for combining the results of different methods (Azadeh et al., 2009).

It is usually desirable to use more than one benchmarking technique for the purpose of methodological cross-checking and to perform diversified analysis. Some argue that a 'preferred model' should be chosen from among them. For example, Haney and Pollitt:

> … there is a question about whether efficiency scores produced by different methods should be combined. Clearly, simply averaging a set of efficiency scores for the same firm (produced for example by DEA, COLS and SFA or different specifications of the same measurement technique) produces a score which itself does not correspond to the result of any one method. It makes more sense to pick the result of one set of estimates, on the basis of the argument that this was the most appropriate method of measuring the efficiency of the sample of firms in question, and consistently use that. (Haney and Pollitt, 2012, pp. 29–30)

A contrasting view is expressed by Agrell and Bogetoft:

> As long as benchmarking scholars cannot clearly rank one method as being superior to another we see no reason the regulator should make that call. It is also not just an 'easy way out' of methodological discussion to apply multiple methods. In fact one can argue that … the simultaneous application of multiple methods puts additional discipline on the model development approach. (2016b, p. 15)

In practice, one model may not entirely dominate the other models, such that an *objective assessor* would inevitably be more convinced by that model compared to the alternatives. Consider the following example: If there are five *objective assessors* and three are convinced that model A is the most reliable, and two are convinced that model B is the most reliable, which is the more appropriate response?

- choose model A over model B; or
- use a weighted average of the results of the two models, with a weight of 3/5 applied to the result of model A, and a weight of 2/5 applied to model B (i.e. weights based on the relative degrees of confidence in the models); or
- some other method of combining the results of models A and B.

Having regard to the evidentiary context of regulatory decisions for which benchmarking is being used, it is not clear that picking one model is always to be preferred, especially if its claims to superiority are only slightly better than some other alternative. In these circumstances, a 'consensus' approach drawing on results from several plausible models has merit. As Farsi et al observed, if there is significant uncertainty in inefficiency estimates, e.g. because there is more than one convincing model that yield different results, this "could have important undesired consequences especially because in many cases the efficiency scores are directly used to reward/punish companies through regulation scheme such as price formulas" (Farsi et al., 2007, p. 13).

8.1 Bayesian model averaging

One perspective on combining models is given by the literature on Bayesian model averaging (BMA). This approach is designed to take account of model uncertainty, which is often ignored, particularly when one model is selected from a number of alternative models, each with some positive likelihood of being the best representation of the data generation process. "This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are" (Hoeting et al., 1999, p. 382).

Once the alternative models have been estimated, the posterior probability distribution of each efficiency score is given by:

$$(7.1) \qquad \Pr(\theta_i|S_n) = \sum_{k=1}^{K} \Pr(\theta_i|M_k, S_n)\Pr(M_k|S_n)$$

where θ_i is the efficiency of firm *i*; S_n is the data sample covering *n* firms; $M_1 \ldots M_K$ are the models considered; and $\Pr(M_k|S_n)$ is the posterior probability of M_k. The probabilities of the different models must add up to one: $\sum_{k=1}^{K} \Pr(M_k|S_n) = 1$. From this formulation the expected value of the efficiency of firm *i* can be expressed as:

$$(7.2) \qquad \mathrm{E}[\theta_i|S_n] = \sum_{k=1}^{K} \hat{\theta}_i \Pr(M_k|S_n)$$

The efficiency measure for firm *i* is a linear combination of the efficiency measures obtained from models $M_1 \ldots M_K$, and the weights are the posterior probabilities for each of those models. The fundamental challenge in applying this method is in the estimation of the probabilities of each model being the 'true' model.

Various algorithms have been developed for this purpose. However, for regression models a simple approximation is available, which uses the Bayesian Information Criterion (BIC), since: $\mathrm{BIC}_k \approx \ln\big(\Pr(M_k|S_n)\big)$ (Elliott and Timmermann, 2016, pp. 339–341). Hence the weights in (7.2) can be approximated by:

$$(7.3) \qquad \Pr(M_k|S_n) \approx \frac{\exp(-0.5.\mathrm{BIC}_k)}{\sum_{k=1}^{K} \exp\,(-0.5.\mathrm{BIC}_k)}$$

A method for estimating information criteria statistics such as Akaike's Information Criterion (AIC) or BIC for DEA models has been developed in the literature on variable selection.[29] Given set of efficiency score estimates obtained from a DEA model, the AIC and BIC statistics for that model can be obtained by regressing those efficiency score estimates against all of the input and output variables that were used in the DEA model, which serve as explanatory variables in the regression model (Li et al., 2017).

Using this approach to estimating BICs for DEA models, and using (7.3) to transform them into weights, the Bayesian model averaging formula (7.2) could be used to average a number of DEA models, or to average a mix of DEA and regression models. For example, this approach could be used to average the efficiency scores of an SFA and a DEA model. Ideally, the models to be averaged will be kept to a minimum, such as the best two or three.

8.2 Minimum Quadratic Loss

Another approach is that of Lavancier and Rochet (2016) which is also based on a linear combination of estimators, and the set of weights (λ's) applied to those estimators are designed to minimize the quadratic loss function: $\mathrm{E}(\hat{\theta}_{\lambda} - \theta)^2$; where $\hat{\theta}_{\lambda}$ is the estimator obtained by combining estimators. This involves finding an optimal set of weights: λ^*. This approach employs a bootstrapping method. It is designed for parametric and semi-parametric applications but can also be used for combining non-parametric estimators. Care is needed to ensure there are not too many estimators being combined, because this makes it difficult to solve for the optimal weights: λ^*. On the other hand, a small a set of estimators could lead to a sub-optimal average estimator $\hat{\theta}_{\lambda}$, but would be easier to estimate. Hence, a good balance is needed between accuracy and ease of obtaining a solution. Because of its reliance on bootstrapping, this method would require a large data sample.

8.3 Discussion

The discussion of BMA suggests that it could feasibly be implemented for combining different DEA models or for combining nonparametric with parametric models. Combining of models using a method such as BMA can be applied to both the estimated efficiency scores as well as the estimated probability distributions that provide the confidence intervals for the efficiency scores. The "primary motivation for BMA is as a way of dealing with model uncertainty" (Elliott and Timmermann, 2016, p. 339).

The DEA and SFA approaches to efficiency measurement each have their own strengths and weaknesses. An approach that combines a preferred DEA model with a preferred SFA model may have merit and is well worth considering.

[29] These two information criteria are closely related: AIC = $-2.LL + 2.p$, where LL is the log likelihood and p is the number of parameters (= number of variables including the dependent variable), and BIC = $-2.LL + \ln N.p$, where N is the number of observations.

9 BENCHMARKING

Benchmarking has been defined as "the process of comparing the performance of one unit against that of 'best practice' units" (Thanassoulis et al., 2008, p. 353). Because data envelopment analysis (DEA) estimates the best practice frontier, based on the firms that are included in the sample, the efficiency scores represent one method of benchmarking a firm's performance against best practice. This is a holistic form of efficiency measurement, and for this reason has been called 'balanced benchmarking' (Zhu, 2015, p. 292). Businesses can be benchmarked against each other in various ways, which is the topic of this chapter.

Comparing a firm's performance against the efficiency frontier is related to target setting for inefficient firms, based on the point on the frontier most relevant to them. This is discussed in section 9.1. Firms can also be compared against other firms, and for inefficient firms the most meaningful comparators are those best practice firms that are similar in terms of input and output mix. Identifying efficient peers against which more detailed comparisons can be made is the subject of section 9.2. Graphical tools that assist in visualising comparisons of the foregoing kinds are discussed in section 9.3.

Another way of comparing the efficiency performances of firms is by ranking them. This can also serve as a means of identifying groups of firms that are more closely related in levels of performance, which may also serve as useful comparators from the perspective of competition by comparison, rather than emulation. Ranking firms and identifying sub-groups is discussed in section 9.4.

Benchmarking can also involve comparing a firm's performance against its past performance, or comparing the *changes* in performance of firms. These types of comparisons rely on the Malmquist productivity index discussed in section 7.2. These time-related types of benchmarking are addressed in section 9.5.

9.1 Target Setting

DEA benchmarking results can be used to formulate targets for input levels that would be required for the DMU to become technically or cost efficient. This section discusses issues in formulating targets. The focus is on the input-oriented efficiency analysis, so the targets would be formulated in terms of inputs. The issue of formulating targets is directly relevant to the use of DEA results in regulation to determine price or revenue paths. The formulation of targets involves identifying the relevant measure of economic efficiency, and also having regard to practical factors that may impede the attainability of targets. Regulatory price or revenue constraints will also need to correspond to efficient and attainable targets.

The radial technical efficiency scores obtained from the basic DEA model represent the factors (for each DMU) by which all inputs could be uniformly contracted in order to reach the efficiency frontier (from an input-oriented perspective). However, radial contraction of inputs need not lead to the optimum economic efficiency for a number of reasons:

(a) As discussed, the point on frontier that an inefficient firm is projected onto need not be a Pareto efficient combination of inputs to produce the given levels of outputs. There may be one or more slacks, which mean that some input(s) could be further reduced while still producing the same output.

(b) Even if (or after) a Pareto efficient point is reached, this is only a technically efficient input combination. It need not be allocatively efficient, and for most DMUs on the frontier (or projected onto the frontier) that will not usually be the case.[30] For a particular DMU, the position on the frontier that is both Pareto technically efficient and allocatively efficient is the input mix that minimises the cost of producing its given set of outputs, given the prevailing input prices.

(c) If there are variable returns to scale, the firm may not be operating at an efficient scale. If not, and if the firm has the ability to change its scale of outputs, then it should be able to achieve greater efficiency by doing so.

Issues (a) and (b) need to be considered when translating DEA model results into target input levels. Issue (c) is not likely to be as important for utility benchmarking, because utilities such as TSOs generally do not have the ability to optimise their scale of operation. They tend to supply the level of demand in the market, which is largely outside their control. However, it could still provide useful information on whether TSOs should be amalgamated or disaggregated.

In principle, two sets of targets can be formulated for the input mix of each DMU. Firstly, the change in input mix needed to achieve Pareto technical efficiency, and secondly, the input mix needed to achieve cost efficiency. These two sets of targets could be quite different targets if there is a great deal of allocative inefficiency. Ultimately the mix of inputs that minimises cost is the most important target, but the nature of these input targets can be more fully appreciated by considering the technical efficiency and allocative efficiency aspects in turn.

For a given firm, DMU$_o$, the individual input targets needed to achieve a (Pareto) technically efficient input mix are given by: $x_{oi}^t = \theta_o^* x_{oi} - s_{oi}^{-*}$, for input *i*. Here x_{oi} is current use of that input, the superscript *t* stands for the technical efficiency target, and θ_o^* is DMU$_o$'s estimated technical efficiency and s_{oi}^{-*} is the estimated slack in its use of input *i* when projected radially onto the frontier. Some two-step methods of quantifying the slacks were explained in section 3.4 (adjusting efficiency scores for slacks was discussed in section 6.3).[31] The set of input targets $(x_{o1}^t \dots x_{om}^t)$ for inputs 1 to *m*, represent the coordinates for the point on the efficient frontier that DMU$_o$ is projected onto (giving priority to radial contraction of inputs in the first instance, and then removing slacks). Target input levels for cost minimisation are obtained directly as the solution values of the DEA cost minimisation program $(x_{o1}^* \dots x_{om}^*)$.

In yardstick regulation frameworks, price or revenue caps are usually formulated with the aim of providing the regulated business with the ability to earn revenue sufficient to meet the

[30] If all DMUs face the same input prices then typically (but not inevitably) there will only be one input combination that is allocatively efficient. However, if they face different sets of input prices, then the cost minimising input combinations will differ between DMUs.

[31] There may be output slacks, even though the model is input-oriented, so there remains an issue of whether they should be taken into account, especially if it is assumed that outputs are outside the control of the firm, and hence it cannot target changes in output.

efficient cost of supply, allowing for the time that may be needed to achieve efficiency. Implicit within these frameworks are targets for inputs, subject to forecasts of demand and input prices. This type of information is likely to be useful to the regulator when setting the regulatory controls, and may also be useful to businesses to translate the revenue or price caps into targets that are directly within their control.

9.2 Efficient Peers & Dominant Firms

Comparisons are commonly made between an inefficient DMU and the most similar of the efficient DMUs. One of the advantages of DEA analysis is the ability to identify efficient peers for each inefficient DMU. The efficient peers of a given DMU$_o$ are firms located on the frontier in the vicinity of the point that DMU$_o$ is projected onto. These are businesses that can produce a similar pattern of outputs, using a similar mix of inputs, but use fewer inputs as a whole. These efficient peers, and no others, determine the efficiency score of an inefficient business. Identifying the peers can help to explain the efficiency findings for a particular DMU because "nonspecialists find the identification of efficient peers especially useful for gaining an intuitive feeling for the comparative efficiency results yielded by DEA" (Thanassoulis et al., 2008, p. 353).

If DMU$_o$ is inefficient, then its efficient peers can be directly identified from the results of the DEA envelopment form program (discussed in section 3.1). In the program run for DMU$_o$ they are the DMUs for which the solution values of the λ's are non-zero. These λ's are weights that define the facet of the production possibility frontier that DMU$_o$ is projected onto, since the efficient input mix at DMU$_o$'s projection point on the frontier is a linear combination of the input mixes of its efficient peers.

To become efficient DMU$_o$ may need to become more like its efficient peers. Therefore, once the efficient peers have been identified, a more detailed comparison can be undertaken, as case studies, between the inefficient TSO and its efficient peers to seek a better understanding of why those businesses are more efficient. The efficient peers can be seen as role models because they have a similar mix of inputs and outputs, and therefore similar operations, and they may operate in similar environments, and what they do differently or better than the inefficient business may shed light on the reasons for its inefficiency. Such comparisons may make use of more detailed operational and financial data as well as key performance indicators (KPIs), and they may be undertaken regularly to enable the inefficient firm to monitor its progress in moving towards a more efficient input-output mix.

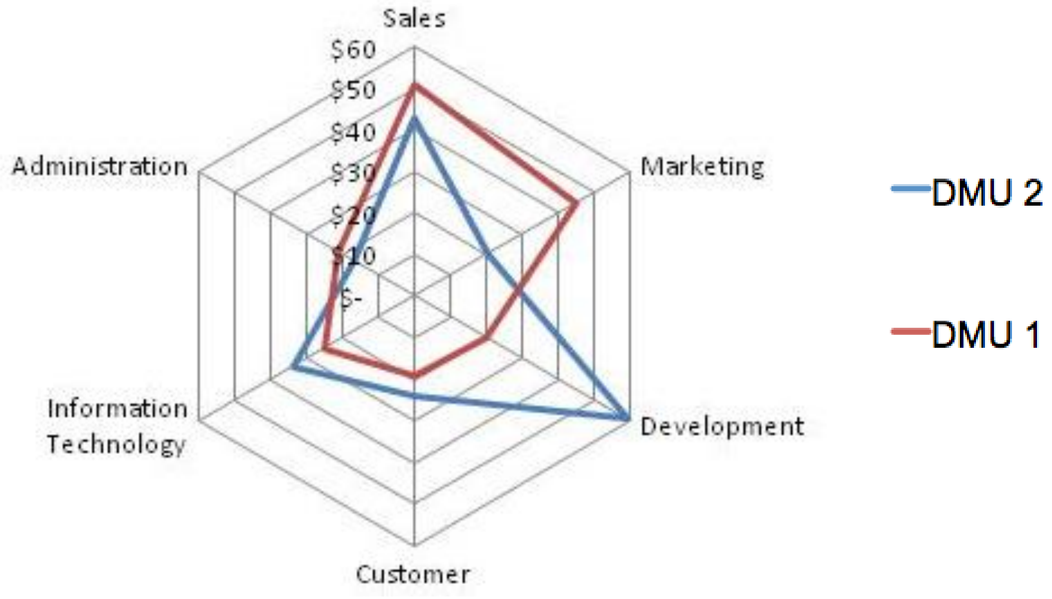
Some of the efficient peers may be of more importance to an inefficient DMU than others. In the free-disposal hull (FDH) methodology, convexity of the frontier is not imposed, and an inefficient DMU is *not* projected onto a point (i.e. a combination of inputs for a given set of outputs) that is a convex combination of efficient peer units. Slacks are far more important in this model, and most of the adjustment to the frontier will typically be associated with removing slacks. Each inefficient firm is projected onto a single actual efficient firm, which is its 'dominant DMU'. This dominant DMU is one of the efficient peers identified in the solution to the basic DEA problem. It might be argued that it is the most important of those efficient peers, and therefore may warrant particular attention.

9.3 Graphical Comparisons

There are various types of descriptive and statistical information that can be informative in relation to the characteristics of the input and output mixes of firms that makes them more or less efficient than others in the sample, or in regard to systematic patterns in efficiency estimates. Examples include two-way graphs of association between efficiency scores and other variables, cross-tabulations of data grouped into intervals or into logical groupings, and correlation measures (eg the Pearson correlation coefficient between continuous variables).

A particularly convenient graphical too is the radar chart (Thanassoulis et al., 2008), an example of which is shown in Figure 4.1. This type of chart can be used to compare one firm against another. Note also that the point on the efficient frontier which an inefficient firm is projected onto can be viewed as a 'virtual DMU' (the inputs and outputs or which are linear combinations of the each of the inputs and outputs of its efficient peers). Hence the radar chart can also be used to compare an inefficient firm against its target inputs and outputs that represent its projection point on the frontier.

Figure 4.1: **Radar Diagram Example**



Source: <https://en.wikipedia.org/wiki/Radar_chart>.

In Figure 4.1, suppose the outputs are Sales and Customers, and the inputs are Administration, Information Technology (IT), Marketing and Development. Positions further away from the centre of the diagram are better for outputs and positions closer to the centre are better for inputs. Two firms are shown in the diagram. The firm shown by the red line (DMU 1) has higher sales but fewer customers than the DMU 2 (shown by the blue line), but overall they are comparable in terms of outputs. DMU 1 uses more Administration inputs but less IT inputs. But where the two firms differ considerably is that the red firm has much higher Marketing inputs, and much lower Development inputs compared to the blue firm. If, for example, DMU 1 was an efficient peer of DMU 2 (or 'virtual DMU' that represented DMU 2's projection point on the efficiency frontier, then the radar diagram would suggest that DMU 2 may want to give consideration to reducing its inputs allocated to development and increasing the inputs applied to marketing. The radar diagram highlights where there may be sub-optimality in the mix of inputs for a given set of outputs.

9.4 Rankings and Subgroups

Ranking the efficiency of firms is an informative descriptive way to present comparisons between the firms in the sample. Changes in rankings over time give a meaningful indication of relative performance, and for any particular firm, their position in the ranking implicitly identifies certain other firms of similar ranking that may be of interest as comparators.

Ranking firms found to be inefficient can simply involve ordering their technical or cost efficiency scores (or both) from highest to lowest. However, efficient firms can't be ranked in the same way because they all have a score equal to one. Methods have been developed to rank the efficient firms based on super-efficiencies.

The concept of super-efficiency was discussed earlier. In a basic DEA model, each firm is included in its own comparator set (ie, the comparator set includes the whole sample) and hence the maximum efficiency score is one. If the DEA model is formulated such that each DMU is excluded from its own comparator set, then some DMUs will have efficiencies greater than one (such as those with scores equal to 1 in the basic DEA model). If the efficiency score for a DMU is greater than one using this method, then it is 'super-efficient'. Since these scores will generally differ between firms, this measure can be used to rank firms that are efficient within the basic DEA model. However, it needs to be noted that the super-efficiency method of ranking is said to have limitations because "the evaluation context changes in the evaluation of each efficient DMU, and the efficient DMUs are not evaluated against the same reference set" (Zhu, 2015, p. 293).

Nonparametric statistics of association can be used to determine the level of association between efficiency rankings derived from different methodologies (e.g. two different models). Exploratory statistics of this type include Spearman's *rho* or Kendall's *tau-b*. They can be used to test the null hypothesis that the efficiency rankings are unrelated to the proposed explanatory variable(s). These statistics are outlined in Box 9.1.

As mentioned, ranking the efficiencies of firms can be useful to identify those firms that are at similar efficiency levels to other firms, and can also be used to identify efficiency quartiles. Identifying the efficiency quartile in which the firm is a member can be useful descriptive information. It can also be particularly useful when making comparisons over time (discussed in chapter 7) because it is of interest to compare the change in a firm's performance against firms that have some kind of similarity. For example, if a firm was previously in the lowest quartile of efficiency scores, it may be informative to compare the change in its productivity to changes in the productivities of the other firms that were in the lowest quartile.

Box 9.1 Nonparametric measures of rank correlation

Spearman's rank correlation coefficient (*rho*) is defined as:

$$r_S = 1 - \frac{6 \sum_k d_k^2}{n(n^2 - 1)}$$

where k refers to observation (DMU) 1 … n, d_k is the difference in the ranking of DMU k between the two analyses (e.g. if ranked 2nd in the first instance and 4th in the second instance then d = 2 – 4 = –2). Like the ordinary correlation coefficient for interval valued variables, r_S can take any value between –1 and +1, and positive values mean that higher ranking in one series tends to go hand-in-hand with a higher ranking in the other series. (There is a specific method for dealing with tied rankings.)

Kendall's rank correlation coefficient (*tau*) is defined as:

$$\tau = \frac{N^{concordant} - N^{discordant}}{n\,(n-1)/2}$$

where $N^{concordant}$ and $N^{discordant}$ are the numbers of concordant and discordant pairs of rankings respectively. Kendall's tau also lies between –1 and +1, and has a similar interpretation to Spearman's *rho.*

There are other ways of developing subgroups for purposes of comparisons. One of these is so-called 'context dependent' DEA. This method involves solving a sequence of DEA models, each with a narrower sample. The motivation of this method is to progressively change the context within which inefficient firms are compared. After the first DEA program is run, and the efficient DMUs have been identified, those efficient DMUs are excluded from the sample, and the DEA program is run again with the smaller sample that includes only those DMUs previously found to be inefficient. A second-level efficiency frontier is then formed by a group of DMUs that are the most efficient of the remaining DMUs. In turn, the second-level efficient DMUs are excluded and the DEA model is computed again, to find a third-level efficiency frontier. This process is continued until there are no more firms in the remaining sample. In this way every firm in the original sample is assigned to either the first, second, third or *n*th efficiency frontier. This is one way of stratifying the sample of DMUs into efficiency level subgroups.

Subgroups developed using 'context dependent' DEA can be used to find DMUs that are in some sense at a similar efficiency level to the DMU of interest, which may be relevant for monitoring changes in performance over time, in terms of whether such changes are similar to those of other firms in the subgroup, or lead to a change in subgroup membership. A final point to make in relation to 'context dependent' DEA is that it may be a useful diagnostic tool when developing a DEA model.

9.5 Comparisons Over Time

Methods of calculating changes in productivity were discussed in section 7.2. It was also shown that it is possible to separate the 'catch-up' effect from the 'frontier shift' effect. TFP growth rates can be compared and ranked. The same can be done with components of TFP growth such as ‘catch-up’. This information can be used to highlight firms whose efficiency has deteriorated and those that have made substantial efficiency gains.

The firms included in the sample can be grouped in various ways and productivity growth rates within each group can be compared. For example, comparisons between firms that have similar characteristics, or that belong to identified sub-groups based on technical or cost efficiency levels. Formal tests of general hypothesis can be carried out, such as whether the extent of catch-up is greater among the least efficient firms, than among firms that are closer to the efficiency frontier.

One comparison of this kind likely to be of particular interest is between the productivity growth rates of firms subject to different regulatory regimes. This type of information may be useful to monitor the effectiveness of the regulatory framework, including whether it is resulting in the efficiency gains that were expected at the time of the last revenue cap determination, and whether there is any correlation between the types of regulation framework and the productivity gains observed.

10 FURTHER TOPICS

This chapter discusses two other topics that are relevant to explaining the results of a benchmarking study. Section 10.1 discusses what should be expected in terms of the substance of benchmarking reports produced for regulatory purposes. Section 10.2 discusses the use of benchmarking frameworks to inform the development by regulated businesses of their own performance management systems such as key performance indictors or balanced scorecards.

10.1 Good Practice Documentation

Part of explaining the outcome of a benchmarking analysis is the standard of documentation that should be provided to the key stakeholders in the experts' report. The key test of good documentation is whether it is sufficient to enable other experts to reproduce the analysis. Because of commercial confidentiality, the information available to regulated businesses, or to their experts, may be much more detailed than that published. Reproducibility is primarily of importance to the experts hired by regulated businesses.

It is standard practice for the benchmarking report to clearly state the questions that the expert was asked to address, and to list the documents, datasets and other materials provided to the expert, or that the expert has been instructed to consider. It should also detail the other documents or data that the expert has relied on.

The Competition Commission, United Kingdom (2009) has issued guidance on the best practice for submissions of technical economic analysis, including modelling and quantitative empirical analysis, which are relevant to the standards for documentation and presentation in expert benchmarking reports. They should be set out clearly and comprehensively and, as far as possible, be understandable to non-economists. They should adhere to three guiding principles:

- *clarity and transparency*: clearly stating the methodology used and assumptions made, their justification, the results and conclusions of the analysis, and the robustness of those results to the assumptions made.
- *completeness*: a complete description of the analysis undertaken, the economic theory employed and the techniques used, with references were relevant to the academic literature. Econometric results should be reproduced with diagnostic tests and robustness checks.
- *replication of results*: data and program files should (subject to confidentiality claims) be made available on request to enable others to reproduce the results. Data sources and all details of data cleaning should be documented.

Where comments have been made on the analysis at a draft stage, then the report should document how those comments have been addressed. If data has been sourced from a survey, then the survey must be fully documented, including the sampling and survey methods, the questionnaire and the raw results.

The European Commission (2010) has also issued guidance on best practices for the submission of economic evidence in competition law matters, which are again relevant here. Among other things it emphasises that:

- the questions of interest should be clearly articulated and the hypothesis to be tested (and the alternative or null hypothesis) should be explicitly formulated;
- the link between the hypothesis being tested and any economic theory should be spelled out;
- the assumptions of the economic model should be consistent with the institutional features and other facts of the industry;
- economic methods and models should be well established in the relevant literature;
- inspection of the data should include summary statistics and graphs and documentation of the data definitions and sources;
- the empirical methods and data should be appropriate to the task at hand, and the results properly interpreted and robust;
- the pros and cons of methodological choices should be explicitly considered;
- counterarguments should be given adequate consideration;
- the plausibility and consistency of the results should be tested against other pieces of quantitative and qualitative evidence;
- the practical relevance of the results should be discussed, and assessed against the relevant economic theory;
- the accuracy or explanatory power of the results should be addressed.

These guidelines suggest that expert benchmarking reports should meet two overall aims. Firstly, they should be sufficiently thorough not only in relation to the documenting data and methodologies in the final analysis, but also with regard to the process of reaching the final analysis, including both the reasoning processes and the quantitative investigation steps. Secondly, the presentation of the study should aim to give the reader an understanding of the key aspects of the analysis and results. For example, by identifying important features of the technology which explain the choices of variables used in the study; aspects of the dataset that have had an important bearing on the results; interpretations of quantitative results in terms of economic theory, and generally to explain and illustrate the results succinctly but effectively.

10.2 DEA and Key Performance Indicators

This section discusses the topic of using economic benchmarking frameworks in conjunction with individual firms' more specific performance frameworks such as *key performance indicators* (KPIs) and *balanced scorecards* (BSCs). The question is whether this can assist firms to operationalize strategies to improve their overall economic efficiency performance under the regulatory benchmarking framework.

Economic benchmarking is used to provide fundamental encompassing measures of business performance in terms of cost efficiency or technical efficiency or both. It sheds light on what is achievable by examining levels of efficiency actually achieved by best practice businesses. The purpose of economic benchmarking within a regulatory framework is to incentivise businesses to achieve cost efficiency, which benefits consumers in the long run. Performance indicators are used by businesses to provide more focus on identifying and implementing efficiency improvements at an activity level and performance against specific business goals. Ideally, performance indicator frameworks provide a sufficiently complete and balanced representation of the activities of the business.

Agrell and Bogetoft observed that there is scope for firms to make greater use of the data collected for regulatory purposes to carry out more detailed analysis for the purpose of developing efficiency improvement strategies:

> Firms that are subject to regulatory benchmarking spend considerable resources collecting and standardizing data, and it is worthwhile to consider what added value this effort may bring to the daily operations of the firms. It can be useful to use the same data to support firm specific learning. (Agrell and Bogetoft, 2016b, p. 32)

KPI or BSC frameworks are often used in businesses for targeting performance improvement but can be less effective without reference to holistic economic performance benchmarking. For example, the limitations of the traditional use of KPIs include not only the difficulty of defining suitable and consistent KPIs, but also the difficulty in prioritising or weighting the performance outcomes when there is a large suite of KPIs to be considered. A common shortcoming is that poor choice of indicators or indicator definitions, or inappropriate balance in the range of KPIs reported, that do not adequately reflect the relevant business objectives, can bias the conduct of the benchmarked businesses and lead to detrimental and inefficient outcomes. There can also be a proliferation of KPIs without a clear logical framework integrating them into a measure of overall performance. Regulatory benchmarking provides a framework within which the relevance and importance of KPIs can be assessed.

DEA is recognised as a useful methodology for holistic economic efficiency benchmarking which provides a valid reference point for specifying balanced performance monitoring frameworks within and between businesses. Furthermore, in a regulatory setting, to be fully effective the KPI or BSC frameworks need to be developed with an understanding of how performance in particular dimensions influences overall economic performance. These observations suggest that business KPI frameworks designed to improve efficiency of particular activities or dimensions of business performance should be complementary to the effectiveness of the regulatory benchmarking framework. These considerations suggest that businesses' own uses for the data may be relevant considerations when formulated reporting requirements.

REFERENCES

ACCC, AER, 2013. Better Economic Regulation of Infrastructure: Country-based Review (Working Paper No. No. 8), ACCC/AER Working Paper Series.

ACCC, AER, 2012. Regulatory practices in other countries - Benchmarking opex and capex in energy networks.

AER, 2017. Draft Decision: APA VTS Australia Gas access arrangement 2018 to 2022; Attachment 7 – Operating expenditure.

AER, 2013. Better Regulation Expenditure Forecast Assessment Guideline for Electricity Transmission.

Agrell, P., Bogetoft, P., 2009. International Benchmarking of Electricity Transmission System Operators, e3Grid Project: Final Report.

Agrell, P.J., Bogetoft, P., 2016a. Endogenous Common Weights as a Collusive Instrument in Frontier-Based Regulation, in: Aparicio, J., Lovell, C.A.K., Pastor, J.T. (Eds.), Advances in Efficiency and Productivity. Springer.

Agrell, P.J., Bogetoft, P., 2016b. Regulatory Benchmarking: Models, Analyses and Applications.

Agrell, P.J., Bogetoft, P., 2014. International benchmarking of electricity transmission system operators, in: 2014 11th International Conference on the European Energy Market (EEM). IEEE, pp. 1–5.

Agrell, P.J., Bogetoft, P., 2010. Endogenous generalized weights under DEA control (Working Paper No. 2010/02). Louvain School of Management Research Institute.

Agrell, P.J., Bogetoft, P., Trinkner, U., 2016. Project E2GAS: Benchmarking European Gas Transmission System Operators (Sumicsid/Swiss Economics).

Agrell, P.J., Farsi, M., Filippini, M., Koller, M., 2013. Unobserved heterogeneous effects in the cost efficiency analysis of electricity distribution systems.

Allen, R., Athanassopoulos, A., Dyson, R., Thanassoulis, E., 1997. Weights restrictions and value judgements in data envelopment analysis: Evolution, development and future directions. Ann. Oper. Res. 73.

Andersen, P., Petersen, N.C., 1993. A procedure for ranking efficient units in data envelopment analysis. Manag. Sci. 39, 1261–1264.

Aparicio, J., Borras, F., Pastor, J.T., Vidal, F., 2015. Measuring and decomposing firmʹs revenue and cost efficiency: The Russell measures revisited. Int. J. Prod. Econ. 165, 19–28. https://doi.org/10.1016/j.ijpe.2015.03.018

Azadeh, A., Ghaderi, S.F., Omrani, H., Eivazy, H., 2009. An integrated DEA–COLS–SFA algorithm for optimization and policy making of electricity distribution units. Energy Policy 37, 2605–2618. https://doi.org/10.1016/j.enpol.2009.02.021

Badunenko, O., Mozharovskyi, P., 2016. Nonparametric frontier analysis using Stata. Stata J. 16, 550–589.

Banker, R.D., Chang, H., 2006. The super-efficiency procedure for outlier identification, not for ranking efficient units. Eur. J. Oper. Res. 175, 1311–1320. https://doi.org/10.1016/j.ejor.2005.06.028

Bogetoft, P., 1997. DEA-based yardstick competition: the optimality of best practice regulation. Ann. Oper. Res. 73, 277–298.

Caves, D.W., Christensen, L.R., Diewert, E., 1982a. Multilateral Comparisons of Output, Input, and Productivity Using Superlative Index Numbers. Econ. J. 92, 73–86.

Caves, D.W., Christensen, L.R., Diewert, W.E., 1982b. The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity. Econometrica 50, 1393. https://doi.org/10.2307/1913388

Clarke, D., 2014. General to specific modelling in Stata. Stata J. 14, 895–908.

Coelli, T., Rao, P., O'Donnell, C., Battese, G., 2005. An Introduction to Efficiency and Productivity Analysis, 2nd ed.

Competition Commission (UK), 2009. Suggested Best Practice for Submissions of Technical Economic Analysis from Parties to the Competition Commission.

Cooper, W.W., Li, S., Seiford, L.M., Zhu, J., 2004. Sensitivity Analysis in DEA, in: Cooper, W.W., Seiford, L.M., Zhu, J. (Eds.), Handbook on Data Envelopment Analysis. Kluwer Academic.

Cullmann, A., 2012. Benchmarking and firm heterogeneity: a latent class analysis for German electricity distribution companies. Empir. Econ. 42, 147–169.

da Silva, A.V., Costa, M.A., Lopes, A.L.M., 2017. Variable selection for electricity transmission Benchmarking: the Brazilian case, in: 2017 International Workshop on Performance Analysis: Theory and Practice. University of Queensland, Brisbane, Australia.

Dai, X., Kuosmanen, T., 2014. Best-practice benchmarking using clustering methods: Application to energy regulation. Omega 42, 179–188. https://doi.org/10.1016/j.omega.2013.05.007

Daraio, C., Simar, L., 2007. Advanced robust and nonparametric methods in efficiency analysis: methodology and applications, Studies in productivity and efficiency. Springer, New York.

Dassler, T., Parker, D., Saal, D.S., 2006. Methods and trends of performance benchmarking in UK utility regulation. Util. Policy 14, 166–174. https://doi.org/10.1016/j.jup.2006.04.001

Economic Insights, 2011. Regulation of Suppliers of Gas Pipeline Services: Gas Productivity, Initial report for the Commission.

Elliott, G., Timmermann, A., 2016. Economic Forecasting. Princeton University Press.

European Commission, 2010. Best Practices for the Submission of Economic Evidence and Data Collection in Cases Concerning the Application of Articles 101 and 102 TFEU and in Merger Cases.

Färe, R., Lovell, C.K., 1978. Measuring the technical efficiency of production. J. Econ. Theory 19, 150–162.

Farsi, M., Fetz, A., Filippini, M., 2007. Benchmarking and regulation in the electricity distribution sector (CEPE Working Paper No. 54). Centre for Energy Policy and Economics, Swiss Federal Institutes of Technology.

Fried, H.O., Lovell, C.K., Schmidt, S.S., 2008. Efficiency and Productivity, in: Fried, H., Lovell, K., Schmidt, S.S. (Eds.), The Measurement of Productive Efficiency and Productivity Growth. Oxford University Press.

Frontier Economics, Consentec, Sumicsid, 2013. E3GRID2012 – European TSO Benchmarking Study.

Gluzmann, P., Panigo, D., 2015. Global search regression: A new automatic model-selection technique for cross-section, time-series, and panel-data regressions. Stata J. 15, 325–

349.

Greene, W., 2007. LIMDEP Version 9.0: Econometric Modeling Guide Vol. 2.

Haney, A.B., Pollitt, M.G., 2012. International Benchmarking of Electricity Transmission by Regulators: Theory and Practice (Working Paper No. EPRG 1226). University of Cambridge, Electricity Policy Research Group.

Hawdon, D., 2003. Efficiency, performance and regulation of the international gas industry—a bootstrap DEA approach. Energy Policy 31, 1167–1178.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 382–401.

Jamasb, T., Pollitt, M., Triebs, T., 2008. Productivity and efficiency of US gas transmission companies: A European regulatory perspective. Energy Policy 36, 3398–3412. https://doi.org/10.1016/j.enpol.2008.05.001

Jenkins, L., Anderson, M., 2003. A multivariate statistical approach to reducing the number of variables in data envelopment analysis. Eur. J. Oper. Res. 147, 51–61.

Johnson, A.L., Kuosmanen, T., 2012. One-stage and two-stage DEA estimation of the effects of contextual variables. Eur. J. Oper. Res. 220, 559–570. https://doi.org/10.1016/j.ejor.2012.01.023

Joskow, P., 2006. Incentive Regulation in Theory and Practice: Electricity Distribution and Transmission Networks.

Kao, C., Hung, H.-T., 2005. Data envelopment analysis with common weights: the compromise solution approach. J. Oper. Res. Soc. 56, 1196–1203. https://doi.org/10.1057/palgrave.jors.2601924

Kaufmann, L., Beardow, M., 2001. External Benchmarks, Benchmarking Methods, and Electricity Distribution Network Regulation: A Critical Evaluation. Pacific Economics Group, Benchmark Economics.

Kittelsen, S., 1993. Stepwise DEA: Choosing variables for measuring technical efficiency in Norwegian electricity distribution.

Kneip, A., Simar, L., Wilson, P.W., 2016. Testing Hypotheses in Nonparametric Models of Production. J. Bus. Econ. Stat. 34, 435–456. https://doi.org/10.1080/07350015.2015.1049747

Kneip, A., Simar, L., Wilson, P.W., 2008. Asymptotics and Consistent Bootstraps for DEA Estimators in Nonparametric Frontier Models. Econom. Theory 24, 1663–1697. https://doi.org/10.1017/S0266466608080651

Kumbhakar, S.C., Parmeter, C.F., Zelenyuk, V., 2017. Stochastic Frontier Analysis: Foundations and Advances.

Kuosmanen, T., 2012. Stochastic semi-nonparametric frontier estimation of electricity distribution networks: Application of the StoNED method in the Finnish regulatory model. Energy Econ. 34, 2189.

Kuosmanen, T., Johnson, A., Saastamoinen, A., 2015. Stochastic Nonparametric Approach to Efficiency Analysis: A Unified Framework, in: Zhu, J. (Ed.), Data Envelopment Analysis: A Handbook of Models and Methods. Springer.

Land, K., Lovell, K., Thore, S., 1993. Chance-constrained Data Envelopment Analysis. Manag. Decis. Econ. 14, 541–554.

Lavancier, F., Rochet, P., 2016. A general procedure to combine estimators. Comput. Stat. Data Anal. 94, 175–192.

Li, Y., Shi, X., Yang, M., Liang, L., 2017. Variable selection in data envelopment analysis via Akaike’s information criteria. Ann. Oper. Res. 253, 453–476.

Llorca, M., Orea, L., Pollitt, M.G., 2014. Using the latent class approach to cluster firms in benchmarking: An application to the US electricity transmission industry. Oper. Res. Perspect. 1, 6–17. https://doi.org/10.1016/j.orp.2014.03.002

Lowry, M.N., Getachew, L., 2009. Statistical benchmarking in utility regulation: Role, standards and methods. Energy Policy 37, 1323–1330. https://doi.org/10.1016/j.enpol.2008.11.027

McDonald, J., 2009. Using least squares and tobit in second stage DEA efficiency analyses. Eur. J. Oper. Res. 197, 792–798.

Olesen, O.B., Petersen, N.C., 2003. Identification and use of efficient faces and facets in DEA. J. Product. Anal. 20, 323–360.

Olesen, O.B., Petersen, N.C., 1995. Chance Constrained Efficiency Evaluation. Manag. Sci. 41, 442–457.

Omrani, H., 2013. Common weights data envelopment analysis with uncertain data: A robust optimization approach. Comput. Ind. Eng. 66, 1163–1170. https://doi.org/10.1016/j.cie.2013.07.023

Orea, L., Jamasb, T., 2014. Identifying efficient regulated firms with unobserved technological heterogeneity: A nested latent class approach to Norwegian electricity distribution networks (Discussion Paper No. 3/2014), Efficiency Series. Departamento de Economía, Universidad de Oviedo.

Orea, L., Kumbhakar, S.C., 2004. Efficiency measurement using a latent class stochastic frontier model. Empir. Econ. 29, 169–183. https://doi.org/10.1007/s00181-003-0184-2

Parmeter, C.F., Racine, J.S., 2013. Smooth Constrained Frontier Analysis, in: Chen, X., Swanson, N.R. (Eds.), Recent Advances and and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr. Springer.

Pastor, J.T., Ruiz, J.L., Sirvent, I., 2002. A statistical test for nested radial DEA models. Oper. Res. 50, 728–735.

Paulun, T., Haubrich, H.-J., Maurer, C., 2008. Calculating the efficiency of electricity and natural gas networks in regulated energy markets, in: Electricity Market, 2008. EEM 2008. 5th International Conference on European. IEEE, pp. 1–5.

Petersen, N.C., 1990. Data envelopment analysis on a relaxed set of assumptions. Manag. Sci. 36, 305–314.

Ray, S.C., 2004. Data Envelopment Analysis: Theory and Techniques for Economics and Operations Research. Cambridge University Press.

Ruggiero, J., 2005. Impact Assessment of Input Omission on DEA. Int. J. Inf. Technol. Decis. Mak. 4, 359–368.

Saati, S., Hatami-Marbini, A., Agrell, P., Tavana, M., 2012. A common set of weight approach using an ideal decision making unit in data envelopment analysis. J. Ind. Manag. Optim. 8, 623.

Santos, S., Amado, C., Rosado, J., 2011. Formative evaluation of electricity distribution utilities using data envelopment analysis. J. Oper. Res. Soc. 62, 1298.

Schmitz, H., Tauchmann, H., 2012. Factor substitution in hospitals: a DEA based approach

(Discussion Paper No. 41/2012). Technische Universität Dortmund SFB 823.

Seiford, L., Zhu, J., 1998. Sensitivity analysis of DEA models for simultaneous changes in all the data. J. Oper. Res. Soc. 49, 1060.

Shleifer, A., 1985. A theory of yardstick competition. RAND J. Econ. 319–327.

Shuttleworth, G., 2005. Benchmarking of electricity networks: Practical problems with its use for regulation. Util. Policy 13, 310–317. https://doi.org/10.1016/j.jup.2005.01.002

Simar, L., 2003. Detecting Outliers in Frontier Models: A Simple Approach. J. Product. Anal. 20, 391–424.

Simar, L., 1996. Aspects of Statistical Analysis in DEA-Type Frontier Models. J. Product. Anal. 7, 177–185.

Simar, L., Wilson, P., 2007. Estimation and inference in two-stage, semi-parametric models of production processes. J. Econom. 136, 31–64.

Simar, L., Wilson, P., 2002. Non-parametric tests of returns to scale. Eur. J. Oper. Res. 139, 115–132.

Simar, L., Wilson, P., 1998. Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models. Manag. Sci. 44, 49–61.

Simar, L., Wilson, P.W., 2001. Testing Restrictions in Nonparametric Efficiency Models. Commun. Stat. - Simul. Comput. 30, 159–184. https://doi.org/10.1081/SAC-100001865

Simar, L., Zelenyuk, V., 2011. Stochastic FDH/DEA estimators for frontier analysis. J. Product. Anal. 36, 1–20. https://doi.org/10.1007/s11123-010-0170-6

Tardiff, T., 2010. Cost Standards for Efficient Competition, in: Crew, M. (Ed.), Expanding Competition in Regulated Industries. Kluwer Academic.

Tauchmann, H., 2011. Partial frontier efficiency analysis for Stata (Discussion Paper No. 25/2011). SFB 823.

Thanassoulis, E., Portela, M., Allen, R., 2004. Incorporating value judgments in DEA, in: Cooper, W., Seiford, L., Zhu, J. (Eds.), Handbook on Data Envelopment Analysis. Kluwer Academic.

Thanassoulis, E., Portela, M., Despic, O., 2008. Data Envelopment Analysis: The Mathematical Programming Approach to Efficiency Analysis, in: Fried, H.O., Lovell, K., Schmidt, S.S. (Eds.), The Measurement of Productive Efficiency and Productivity Growth. Oxford University Press.

Tone, K., 2004. Malmquist Productivity Index: Efficiency Change Over Time, in: Cooper, W., Seiford, L., Zhu, J. (Eds.), Handbook on Data Envelopment Analysis. Kluwer Academic.

von Geymueller, P., 2009. Static versus dynamic DEA in electricity regulation: the case of US transmission system operators. Cent. Eur. J. Oper. Res. 17, 397–413.

Zhu, J., 2015. DEA Based Benchmarking Models, in: Zhu, J. (Ed.), Data Envelopment Analysis: A Handbook of Models and Methods. Springer.

Zhu, J., 2003. Quantitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spreadsheets and DEA Excel Solver. Springer US, Boston, MA.

Zieschang, K., 1984. An Extended Farrell Technical Efficiency Measure. J. Econ. Theory 33, 387–396.