Regulating Beyond Price

Integrated Price-Quality Regulation for Electricity Distribution Networks

Virendra Shailesh AJODHIA

Regulating Beyond Price
Integrated Price-Quality Regulation for Electricity Distribution Networks

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof.dr.ir. J.T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op dinsdag 24 januari 2006 om 13:00 uur

door

Virendra Shailesh AJODHIA

elektrotechnisch ingenieur
geboren te Amsterdam

Dit proefschrift is goedgekeurd door de promotor:
Prof.dr.ir. M.P.C. Weijnen

Samenstelling promotiecommissie:
Rector Magnificus, voorzitter
Prof.dr.ir. M.P.C. Weijnen, Technische Universiteit Delft, promotor
Dr.ir. R.A. Hakvoort, Technische Universiteit Delft, toegevoegd promotor
Prof. J.M. Glachant, University Paris XI
Prof.ir. W.L. Kling, Technische Universiteit Delft
Prof. I.J. Peréz-Arriaga, Universidad Pontificia de Commillas de Madrid
Prof. P. Ranci, Università Cattolica del Sacro Cuore

Cover design: Studio TMC
ISBN-10: 9090196528
ISBN-13: 9789090196527

Preface

I should start by thanking Margot Weijnen who provided me with the opportunity to do this PhD project. I am very indebted to Rudi Hakvoort who as my supervisor guided me in my journey towards academic enlightenment. My thinking has certainly changed since 2001 when I started this thesis. For his enduring support from 1992 on till today, I am thankful to Wil Kling. Also to my colleagues at KEMA but in particular my two brothers in arms – Konstantin Petrov and Gian Carlo Scarsi from the Bonn office – for the great times we had travelling and working together around the globe. The work presented here has substantially benefited from these experiences and the many papers we wrote together. And then there is Misja Mikkers, my roommate at DTE; our many heated theoretical debates sometimes left innocent bystanders in a confused state of mind. Also, Victoria Shestalova and Martijn van Gemert contributed to the many discussions we had on regulation and benchmarking. And from the good old DTE times, there is always Cees van Gent who gave me the possibility to combine my research with the alchemy of applied regulation. With my colleagues at Delft University – Francois Boisseleau, Harish Goel, Hamilcar Knops, and of course Laurens de Vries – I had numerous discussions and the opportunity to withdraw from daily regulatory chaos to analyse the world comfortably as an academic. A special gracias to Luis Olmos for the summer of 2004 when we developed the NST. And although I don't like Italian food, I must thank Roberto Malaman and Luca Lo Schiavo for giving me the great opportunity to spend three weeks in Italy to study the quality regulatory system. Two people at KEMA had a special role in this PhD whom I should thank. Harold Dijk made me realise that a PhD in the end is also just a project that needs to be managed. And Arnold Bovy perhaps doesn't realise it but he was the one who inspired me into pursuing a PhD when I had just joined KEMA. I also say a warm thank you to all my clients, colleagues, and friends from utilities and regulators all over the world who made me understand and appreciate the true complexities of regulation. Finally, I thank my parents Jules and Lucia Ajodhia and sisters Kiran and Vidya for their love. Last but certainly not least, I would never have made it without my wife Shashi and our daughter Juhi, whose everlasting love and support provided continuous inspiration to complete this thesis.

Virendra Ajodhia

Paramaribo, 25 November 2005

Table of Contents

1

Introduction

1.1 Background

1.1.1 Power Sector Reform

There is a worldwide trend of countries reforming their power sectors: Liberalisation and privatisation have been introduced and a new approach is taken to the regulation of the remaining network monopolies. Generally, the main objectives of power sector reform have been to improve efficiency and quality levels (Newbery 1999). Economic theory predicts that, in general, firms operating under competition perform better than those under monopoly. Competition leads to higher economic efficiency as producers will continuously seek to increase profits by operating more efficiently and by adapting the quality of their products to consumers' demand. This in turn leads to lower prices and higher quality – ultimately at the benefit of consumers.[1]

Competition, however, is not feasible in all segments of the power sector; transmission and distribution networks remain natural monopolies. Effectively, power sector reform divides the industry into two playing fields: Market and Monopoly. The market part contains those functions that have been liberalised; this typically includes generation, wholesale and retailing as well as some ancillary functions such as metering and billing. The monopoly part consists of the transmission and distribution networks and their related functions such as system control and balancing. In the reformed power sector, policy makers face issues revolving around, roughly speaking, three main areas: There where competition has already been introduced (the market), there where competition is not yet feasible (the network monopolies), and there where interactions occur between market and monopoly.

Regarding the market, it is important to make sure that liberalisation is effective i.e. competition actually develops. To promote competition, a suitable legal framework has to be set in place as well as supporting institutions such as independent system operators and power exchanges. To maintain fair competition, new entrants should be accommodated and issues like market concentration and collusion should be dealt with appropriately. In addition, incentive mechanisms need to be put in place to assure availability of generation capacity in order to avoid any security of supply problems. Regarding the distribution and transmission networks, these remain natural monopolies and continue to be economically regulated. Here, not competition, but regulation is expected to generate benefits similar to those in the market: Lower prices and better quality levels. Finally, the interface between market and monopoly also needs to be properly organised. Responsibilities and duties of market parties and monopoly network operators need to be defined clearly and rules have to be set in place to organise the way these two interact.

This thesis deals with the issues surrounding the monopoly business: The economic regulation of network monopolies.[2] At this time, the natural monopoly character of the networks implies that the transmission and distribution functions cannot yet be supplied in competition. That does not refrain policy makers, however, from pursuing higher economic efficiency for this part of the power sector. Liberalisation of the power generation, wholesale and retail markets is often accompanied by a parallel introduction of strict price controls for the network monopolies. These new price controls – known as price-cap regulation – provide a strong incentive to decrease costs as they unlink prices from actual costs. Under a price-cap, cost savings translate into higher profits. There is a growing concern that the drive towards network cost savings may result in problems at the quality front if it cannot be excluded that the firm attains (part of the) cost savings through an - undesired - reduction in quality. This makes the inclusion of quality into the regulatory framework an important aspect of price-cap systems.

1.1.2 The Network Quality Problem

The list of countries that have made the move to price-cap regulation is a long one.[3] At the same time, there is a growing concern as well as empirical evidence that this change – or more specially the stronger focus on cost reduction – may create adverse effects on quality. In their seminal contributions, Michael A. Spence (1975) and Eytan Sheshinski (1976) already predicted and discussed the potential problems of quality under price-cap regulation. Subsequent authors have further studied the relations between price regulation and quality (see also chapter two). For electricity distribution, the literature suggests that a move to stricter price controls, such as those based on price-cap regulation, leads to reduced network quality. Not only the literature acknowledges the quality problem under price-caps, empirical evidence now confirms the need for explicit quality provisions under systems of price-cap regulation. In a recent empirical study, Ter-Martirosyan (2003) conducts a detailed study of reliability performance in 23 States of the US. The main finding of her analysis is that States, which moved to stricter forms of price control, indeed experienced a decrease in network reliability levels – as reflected by an increase in the interruption frequency and average interruption duration. This finding supports the theoretical concern that price controls based on price-cap regulation have unfavourable effects on network quality.[4] Furthermore, and more importantly, she found that these unfavourable effects were diminished when the regulator had implemented quality regulation controls in addition to the price-cap. These empirical findings confirm the theoretical concern for quality provisions under price-cap regulation.

Quality regulation is not a new topic. Concern about quality under monopoly has always been there, but only after the introduction of price-cap regulation, attention for it has become widespread. The main reason for this is the perceived danger of degrading quality under price-cap regulation. This problem was already acknowledged by Professor Stephen Littlechild (Littlechild 1983) in his pioneering contribution that paved the way for the introduction of price-cap regulation in Britain as well as other countries. More recently, the Council of European Energy Regulators noticed (CEER 2003):

> *"…simple price-cap regimes could incentivise a regulated firm to reduce its quality of supply by cutting investments, maintenance, or personnel with the aim of increasing its profits."*

In the US, the National Association of Regulatory Utility Commissioners points out that (NARUC 1997):

> *"[Under price-cap regulation]…unfettered incentives to reduce costs could result in unacceptable declines in service quality."*

Having established that there is a need to counter the threat of network quality degradation under price-cap regulation, the question then is what options exist for regulators to choose from and how well these options perform.

1.2 Research Objective and Questions

1.2.1 Scope and Definitions

Before proceeding further, it is helpful to first define the scope of this research. Along with this, the definitions of regulation, the distribution network, and quality are also provided.

Competition is generally preferable to monopoly because it is believed to generate better and socially more desirable outcomes. The exploitation of monopoly power leads, among others, to economic inefficiency as well as suboptimal quality levels. If monopoly is unavoidable, for example in the case of a natural monopoly, then regulation is usually applied to counteract such adverse outcomes. Regulators are assigned the task to attain objectives that are beneficial for society and these typically include the promotion of high economic efficiency and adequate levels of quality. Regulation, in the context of this thesis, is defined as the imposition of methods, rules, systems, etc. that aim to counteract perverse monopoly outcomes and bring price and quality towards socially desirable levels.

It is important to point out the difference between regulation (price and quality control) and pricing (tariff design). Regulation covers the mechanisms through which the regulator determines the allowed income for the firm. Pricing or tariff design is concerned with the process of determining the structure of these tariffs.[5] Price control and tariff design are two related but still separate issues. Price control deals with the determination of the allowed income to the firm. Pricing relates to the issue of designing a bundle of tariffs so that these can together generate the allowed income for the firm as determined by the regulator (price control). Typically, tariffs are differentiated by consumer type (e.g. households, small industries, large industries) and consist of various elements (e.g. fixed charge, energy charge, capacity charge). In this thesis, the focus will be on price control; issues related to pricing will not be considered.[6]

This thesis deals with price and quality regulatory issues related to the network and in particular to the distribution network. The physical electricity system can be divided into three main segments: Generation, Transmission, and Distribution. Electricity is produced in generating plants and these typically feed into the transmission network.[7] The transmission network effectively connects all generation plants and acts as an interface to the distribution network. At the terminals of the transmission network, voltage is transformed to lower levels; this is where the distribution network starts. The distribution network's function is to take the electricity from these terminal points to the final consumers. Thus, distribution networks are characterised by a uni-directional flow of electricity from these terminal points to the final consumers. In contrast, transmission networks feature bi-directional flows i.e. in both directions. The distribution network can be divided into the medium and low voltage network (MV and LV).[8] The MV network distributes the electricity from the terminals of the transmission system to the smaller MV/LV transformer stations. From here on, the LV network distributes the electricity further to the final consumers. In practice, most of the interruptions in the electricity supply have their source in the MV network.[9] At the same time, the cost of the MV network forms a considerable part of the total costs of electricity distribution firms.[10]

The term 'quality' itself is open to many different interpretations. According to the Oxford Advanced Learner's Dictionary, quality of a product is defined as the "degree of goodness or worth" of that product. For electricity distribution, it is common to make a distinction between three different quality dimensions: Voltage quality, commercial quality, and reliability (CEER

2001). Firstly, voltage quality, sometimes called power quality, covers a variety of disturbances in a power system. It is mainly determined by the physical quality of the voltage waveform. The relevant technical phenomena are variations in frequency, fluctuations in voltage magnitude, short-duration voltage variations (dips, swells, and short interruptions), long-duration voltage variations (over- or under-voltages), transients (temporarily transient over-voltages), and waveform distortion (Dugan et al. 1996). Secondly, commercial quality is related to individual agreements between the distribution firm and their consumers. Examples of such agreements are the conditions for (re)connection of new consumers, installation of measuring equipment, regular transactions such as billing and meter readings and sporadic transactions such as responding to problems and complaints.

The third quality dimension is reliability, which is a measure for the ability of the network to continuously meet the demand from consumers. Network reliability can be divided into two main elements namely adequacy and security (Kling 1994). Adequacy relates to the availability of sufficient network capacity to guarantee supply of electricity to consumers in the longer run. That is, no interruptions occur under normal operating and demand conditions. Security relates to the ability of the network to – given that it is adequately designed – withstand disturbances i.e. consumers do not experience an interruption in the electricity service.[11]

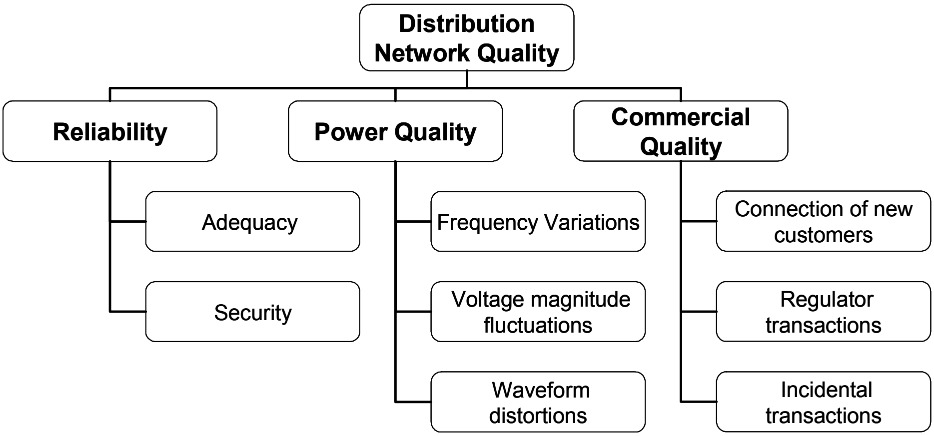


Figure 1-1. Electricity distribution quality dimensions.

From the three quality dimensions, reliability is by far the most important quality feature in electricity distribution. The reason for this is that it is generally considered the core value of electricity service provision. Any service interruption temporarily ceases the provision of electricity and therewith directly affects consumers. This thesis therefore focuses on the reliability dimension of electricity distribution quality. This does not imply that the two other quality dimensions are unimportant. Power quality and commercial quality, however, are of little value if not first and foremost the reliability of service provision is ensured.

Distributors usually measure reliability by collecting information on the number and duration of interruptions and presenting this information through indicators. Reliability indicators represent average values of a particular reliability characteristic for an entire system, operating region, substation territory, or feeder. There are a large number of indicators available for measuring the reliability in distribution networks. Eventually, however, all these indicators in one way or the other can be traced back to the number and duration of interruptions experienced by consumers in a predefined period (usually, one year). The way these two sources of information – the number and duration of interruptions – are weighted and normalised defines the type of reliability indicator; each combination of weighting and normalisation factor in principle leads to a different type of indicator.

In practice, reliability indicators tend to be clustered around three types of factors: Firstly, "customer-based" indicators that relate to the number of consumers affected are the ones most frequently used by firms and regulators. The second class of indicators relates to transformer capacity and is denoted as "load-based" indicators. Finally, the third class of indicators is based on the amount of energy not supplied and hence denoted as "energy-based" indicators. A more detailed overview of the different reliability indicators is provided in Annex I.

1.2.2 Research Objectives

Information – or rather, the lack of information – is a central theme in modern regulation theory. The importance of information is in particular highlighted by the influential work of Laffont and Tirole (1993). In what they call the "new regulatory economics", the regulator's inferior informational position is presented as the primary cause of his inability to achieve his objectives fully. In principle, when the regulator would be perfectly informed, he would also be able to set in place the optimal regulatory policy in order to achieve the societal objectives. However, lack of information prevents him from doing so. It is not likely that the regulator will ever have perfect information, but generally, more and better information leads to more effective regulation and consequently better outcomes. This thesis can be considered an effort to overcome the regulatory informational disadvantage. In particular, it considers the issue of quality under forms of price-cap regulation.

Price-caps generate strong incentives to reduce costs. Ill-designed price-caps may generate a perverse incentive to reduce costs through undesired quality reductions. In principle, it is not in society's interest if the benefits from price-cap regulation come at the expense of quality degradation. Consumers do not only attach value to low prices, but also appreciate decent quality levels. Lower prices associated with quality deterioration are not likely to leave the consumer better off. The challenge then is how to avoid adverse quality degradation i.e. how to make sure that both price and quality are at appropriate levels. This raises the question of how quality can be integrated into the price-cap system. This leads to the formulation of the overall objective of this thesis, which is:

TO DEVELOP AN INTEGRATED APPROACH FOR OPTIMAL PRICE-QUALITY REGULATION OF ELECTRICITY DISTRIBUTION NETWORKS UNDER A SYSTEM OF PRICE-CAPS.

This thesis is about integrated price and quality regulation of electricity distribution networks; price and quality are not studied in isolation but in an integrated fashion. There is a substantial body of literature on issues related to price regulation. Quality issues, however, are far less addressed in the literature. If quality aspects are considered at all, this is usually done from an isolated point of view, i.e. without considering the potential impact on price issues. Similarly, discussions related to price regulation tend to ignore quality issues or, if they do, only provide limited analysis of any quality impacts. A truly integrated analysis of price and quality regulation for electricity distribution networks is still lacking. This thesis aims to fill this knowledge gap.

1.2.3 Research Questions

In order to achieve the above research objective, a number of research questions should be answered. The first question relates to what the optimal outcome in terms of price and quality is. This task is thought to be delegated to the regulator who acts in the best interests of society. His problem is to identify and attain the optimal price and quality pair. This leads to the definition of the first research question:

RQ1. WHAT IS AN OPTIMAL PRICE AND QUALITY LEVEL?

Once it is clear what to aim at, the next step is to consider how to achieve this target. The regulatory approach is defined as the whole of strategies, instruments, methods, etc. applied by the regulator in order to achieve the given price and quality objectives. It is likely that each approach will perform differently in this respect. The next research question aims to explore this issue:

RQ2. WHICH APPROACHES EXIST TO REGULATE PRICE AND QUALITY AND HOW EFFECTIVE ARE THEY?

The first two research questions provide a better understanding of the regulatory objectives and possible approaches to achieve them. They will also reveal the strong and weak points of the different regulatory approaches. The final research question takes this analysis one step further and explores ways for improvement. That is, the scope for increasing the effectiveness of the given regulatory approach:

RQ3. WHICH MEASURES CAN BE DESIGNED/DEVELOPED TO IMPROVE THE EFFECTIVENESS OF THESE REGULATORY APPROACHES?

1.3 Research Approach

1.3.1 Multidisciplinary Nature

Regulation of network monopolies has traditionally been treated as an economic problem. The presence of monopoly leads to both low efficiency and suboptimal quality. The regulator's task is, simply stated, to prescribe a price and a quality level so that economic efficiency is achieved, including an optimal level of quality. In carrying out this task, the regulator is severely hindered by lack of information. The better the regulator is able to overcome this informational barrier, the better he can achieve his objectives. This, as is now generally accepted, can be achieved by providing the regulated firm with the appropriate incentives (Laffont and Tirole 1993).

At the same time, the electricity network – which is the subject of regulation – is a technically complex system. The design of an appropriate incentive system to steer the operation and management of the distribution networks towards socially optimal network performance requires a solid understanding of the networks' technical complexities. Integration of these engineering complexities into the regulatory system may help to improve the design of economic incentives and consequently lead to more effective regulatory approaches, in particular with a view to quality regulation.

This thesis combines the above disciplines i.e. those of economics and engineering. In particular, this thesis is a blend of regulatory economics and electrical engineering. This mix of social science and engineering disciplines reflects the multi-dimensional nature of the system to be regulated: On the one side, the electricity distribution network is a complex physical network. On the other side, it can be considered as a complex social network, involving a multitude of public and private actors pursuing their specific public and private goals, respectively. Where conflicts of interests arise, an appropriate legislative and regulatory framework must be in place to ensure that competing public interests are reconciled and/or that public and private interests are reconciled. The actor system, which includes the regulator, governs the development of the physical system and vice versa. The combination of disciplines in this thesis is believed crucial to untangling the combination of economic and technical complexities of the regulatory challenge at hand: The integrated price-quality regulation of electricity distribution networks.

1.3.2 Research Methods

The thesis can be divided into two main parts namely an analytical part and a design part. The analytical part corresponds to the first two research questions while the design part corresponds to the third research question.

In the analytical part, the underlying concepts and approaches for regulation are explored and evaluated. In particular, the need for integrated price-quality regulation and alternatives for doing so are developed and evaluated. The main output of this part is identification of two approaches for integrated price-quality regulation and an analysis of the associated complexities involved in applying them.

Literature review (concentrated in chapter two) plays an important role in the analytical part. This review includes both the academic and non-academic literature. The academic literature consists mainly of publications on the economic and engineering issues of the regulatory problem. The non-academic publications include regulatory publications such as electricity laws, consultation documents, and regulatory decisions. In addition to literature review, the analytical part also makes use of field research. This consists of interviews with different regulatory specialists both in a formal setting as on an informal basis.

The analytical part provides the theoretical framework for integrated price-quality regulation. The design part aims to bridge the gap between the conceptual and applied stages of integrated regulation. This gap comes in the form a lack of regulatory tools to properly measure combined price-quality performance and therefore effectively configure an integrated regulatory incentive scheme. During the design phase, two methodologies for integrated price-quality benchmarking are developed. These methods are empirically verified by applying them to real-world data and through conducting sensitivity analysis.

1.4 Outline of the Thesis

The structure of the thesis is presented in Figure 1-2. In line with the research approach highlighted above, two main parts – apart from the introduction and conclusions – can be identified. The first part, which includes chapter two till four, is the analytical part of the thesis. The second part of the thesis, which includes chapters five and six, forms the design part of the thesis.

Chapter two performs a literature review of the issues surrounding price and quality regulation of natural monopolies The two main systems to regulate monopolies, rate-of-return regulation and price-cap regulation, are studied and particular attention is paid to their effects on price (essentially, productivity) and quality (essentially, reliability).

Chapter three develops a taxonomy of price-cap approaches and evaluates these on the basis of their quality impact. In particular, the role of benchmarking in the determination of the X-factor – the main element of a price-cap – is highlighted. Two main regulatory approaches emerge: “total costs” (totex) and “building blocks”.

Chapter four considers the issue of optimal quality regulation. Different quality controls – as observed in practice – are evaluated in the light of their incentive for optimal quality. Also, the main problems involved in developing an optimal quality incentive scheme are studied.

Chapter five first deals with integrated price-quality benchmarking; this corresponds to the totex approach. In this chapter, traditional cost benchmarking models based on Data Envelopment Analysis (DEA) are extended to include quality. The newly proposed benchmarking method is applied to an international data set of distribution firms and evaluated with respect to the regulator's informational problems identified in the first part of the thesis.

Chapter six considers integrated regulation under the building block approach. A new methodology for integrated price-quality evaluation of investment proposals is developed and implemented into a software tool (Network Simulation Tool - NST). The chapter presents the underlying methodology of the NST and tests the model by applying it to a representative case of distribution network planning.

Finally, chapter seven presents the conclusions and makes recommendations on further improvement and extensions of this research.

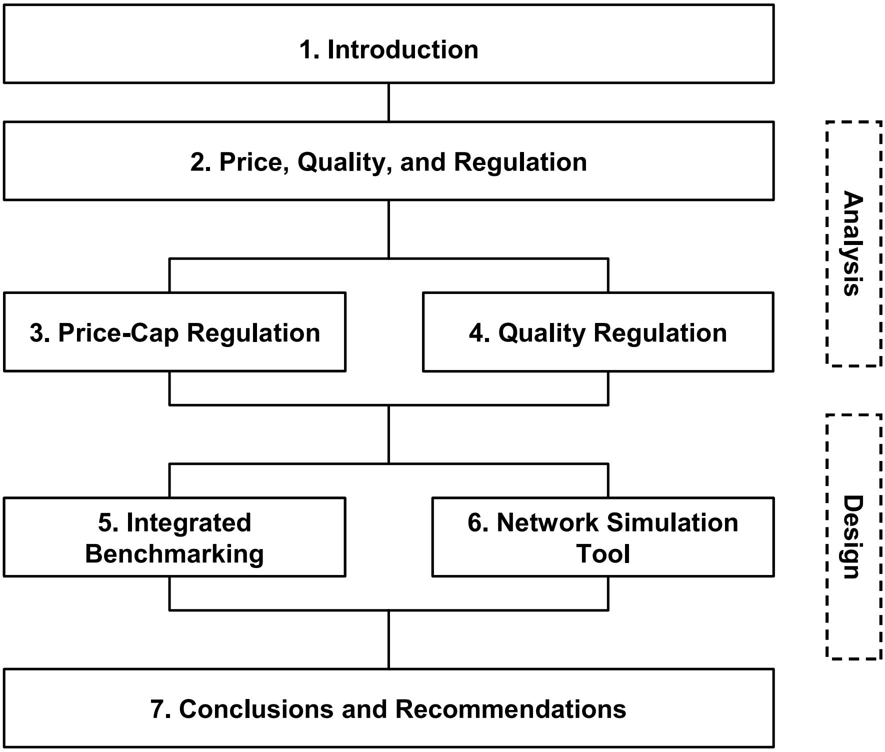


Figure 1-2. Schematic representation of the contents of the thesis.

Notes

[1] The term "consumer", which is used throughout this thesis, should be interpreted as the end-users of the distribution network system and includes both household, commercial, industrial and other types of end-users.

[2] Issues related to the organisation of the market and market-monopoly interfaces are studied, among others, by researchers at Delft University. See for example Boisseleau (2004) for an analysis of the role of power markets in liberalised energy markets. Knops (2005) studies the legal aspects of power sector reform. De Vries (2004) considers the problems of generation adequacy and coordination between market and monopoly.

[3] Jamasb and Pollit (2000) mention Austria, Denmark, Finland, the Netherlands, Ontario (Canada), England & Wales, Northern Ireland, Sweden, Ireland, Hungary, Spain, Japan, California (USA), Australia (Victoria, Queensland, Tasmania and New South Wales), Italy, Brazil, Colombia, India, Chile and Norway. Sappington et al. (2001) identify sixteen states in the US that have moved to some form of performance based regulation. Other countries that have recently implemented forms of price-cap systems include Belize, Jamaica, Romania, Singapore, Slovenia, and Trinidad and Tobago.

[4] It may be that a quality decrease is favourable if initially quality was too high.

[5] Price control and pricing are sometimes compared as respectively "baking the pie" and "slicing the pie" (Peréz-Arriaga 2004).

[6] Optimal pricing includes the problems of how to use tariffs to promote economic optimal allocation as well as other objectives such as protecting low-income consumers, promoting demand side initiatives, etc. The main idea is to start with tariffs reflecting marginal costs and then adjusting these to meet the break-even requirement while at the same time taking into account the other economic or political considerations. Comprehensive treatments on pricing applications can be found in Munasinghe and Warford (1982), Munasinghe (1990) and Train (1991). A discussion of application of marginal cost pricing by EDF in France is provided by Chick (2002).

[7] There may also be generation within the distribution network (dispersed generation) – this is not within the scope of this thesis.

[8] In Europe, MV and LV voltage levels are typically 10 kV and 0.4 kV respectively. This is the phase-to-phase voltage which is a factor $\sqrt{3}$ larger than the phase-neutral voltage.

[9] For example, in the Netherlands, interruptions due to MV outages contributed to 67 percent of cumulated interruption time during the period 1976-2000 Energiened (2001, p. 13).

[10] For example, in the Netherlands, MV assets make up about 40 percent of the total (HV+MV+LV) network asset base of electricity distribution firms. Source: Own calculations from Annual Reports of Dutch distribution firms.

[11] It is important to note the difference between the terms interruption and outage. An interruption is a cessation of electricity service. An outage is the failure of one or more components of the electricity system to do their job, either because of actual equipment failure, damage by weather or other causes, or because they have been switched out of service, either deliberately or by mistake because of failure of control equipment. Thus, an interruption is caused by one or more outages, but an outage does not necessarily lead to an interruption (Willis 1997, p. 76).

2

Price, Quality, and Regulation

2.1 Introduction

2.1.1 Background

Price and quality (regulation) are two related issues and this thesis aims to study them in an integrated fashion. However, before doing so, it is useful to study each of these two issues in isolation and in relation to each other. This can contribute to the development of effective joint price and quality regulatory systems. Chapters three and four look more closely at the options to regulate price and quality and to integrate these two into a single regulatory framework. Before doing so however, this chapter first reviews the theory underlying price, quality, and monopoly regulation. Here, price and quality outcomes under monopoly and the role of regulation are analysed. Also, the main regulatory systems and their impact on quality are reviewed.

2.1.2 Chapter Outline

Section two starts by looking at the main source of the regulatory problem - the presence of natural monopoly. It then considers the effects of monopoly on price and on quality. As will become clear, monopoly leads to low economic efficiency and suboptimal quality levels. Section three presents regulation as a method to counteract these unfavourable outcomes. Also, some of the critiques on regulation are considered. Sections four and five discuss the two main systems for regulation namely rate-of-return regulation and price-cap regulation, respectively. These two systems are evaluated in the light of the incentives they provide for cost efficiency and quality. In particular, the potentially adverse effects on quality under price-cap regulation are considered.

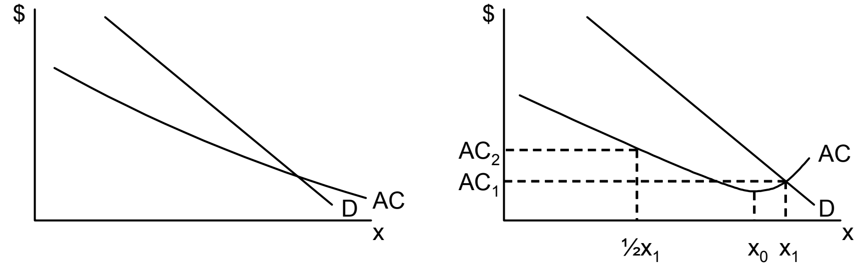
2.2 Price and Quality under Monopoly

2.2.1 Natural Monopoly

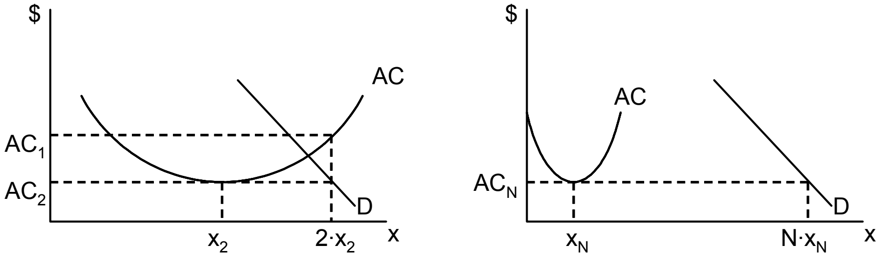
It is well known that competition stimulates economic efficiency and that this generally leads to increased welfare at the benefit of society. Sometimes however, competition is not feasible. Such a situation applies in the case of natural monopoly. Natural monopolies arise if duplication of an infrastructure or service is uneconomic, i.e. the character of the technology and demand dictate that the service is cheaper if the market is served by a single firm rather than by competing firms. The underlying source of this problem is subadditivity of costs (Train 1991). Assume that one firm produces a given level of output x and that there are two other firms producing x' and x'' such that $x'+x''=x$. Then, subadditivity implies that $c(x < c(x')+c(x'')$ in which $c(x)$ stands for the costs to produce the given output x. More generally, a natural monopoly arises if the condition $c(\sum x_i) < \sum c(x_i)$ applies.

The main sources of the existence of subadditivity of costs are economies of scale and economies of scope. Economies of scale imply that average costs fall with increasing output. The most prevalent source of economies of scale is fixed costs, costs that are incurred irrespective of the level of output. Whether or not a natural monopoly is sustainable, however, depends on the range of economies of scale relative to the market demand. In the standard situation, as shown in Figure 2-1(a), average costs decrease over all output levels and clearly there will be a natural monopoly. However, if average costs decrease, but not with *all* levels of output, for example a “U-shaped“ average costs curve as in Figure 2-1(b), and demand is located in the increasing section of the average costs curve, but not too far away from the bottom of the U-shape, still a natural monopoly situation arises. If average costs decrease and then increase, a natural monopoly will exist until demand splits the market equally between firms. Then, two firms will both produce an output x_2 at average costs of AC_2. A single firm producing all output $2\cdot x_2$ would incur higher average costs. This border case, as depicted in Figure 2-1(c), is referred to as a natural duopoly. Competition finally occurs when economies of scale are exhausted at a level of output that is small, compared to market demand.

Next to economies of scale, economies of scope also lead to subadditivity of costs. Economies of scope arise if a given quantity of each of two or more goods can be produced by one firm at lower costs than if each good were produced separately. Assume that the costs to produce two goods of quantity x_A and x_B are given by $f(x_A,0)$ and $f(0,x_B)$ respectively. Combined production cost is given by $f(x_A,x_B)$. Economies of scope exist if $f(x_A,x_B) < f(x_A,0)+f(0,x_B)$ i.e. when it is cheaper for a single firm to produce both products.



(a) Natural Monopoly **(b) Natural Monopoly**



(c) Natural Duopoly **(d) Competitive Market**

Figure 2-1. The relation between average costs and demand (Train 1991). *AC*= average cost, *D*=demand, *x*=output.

Networks – including electricity distribution networks – provide a clear example of natural monopolies (Newbery 1999, p. 27). Here, large scale economies exist as a result of the large portion of fixed costs. By their very nature, as can be observed in Table 2-1, electricity networks are very capital intensive. Electricity distribution networks also comply with the natural monopoly definition if the more elaborate list of characteristics of natural monopolies, provided by Farrer (1902)[1], is taken into account:

- Economies of scale;
- Capital-intensity;
- Non-storability with fluctuating demand;
- Locational specificity generating location rents;
- Producing necessities or essential for the community;
- Involving direct connections to consumers.

Electricity networks fit Farrer's list perfectly (Newbery 1999, p. 28): The network is an obvious case where duplication raises the total cost of supplying market and hence meets the modern definition of natural monopoly. If demand fluctuates, and the product or service cannot be stored, then capacity will need to be sized to peak demand, or demand rationed. Locational advantage suggests that one firm will obtain at least a local monopoly, and different firms may enter to exploit different locations. Finally, the combination of necessity and direct connection implies large market power and the risk of market power abuse by the firm, so that regulation and/or public ownership is politically inevitable.

Table 2-1. Capital intensity of energy firms in the Netherlands, measured as depreciation per employee in 2002. Source: CBS Statline (www.cbs.nl) and own calculations.

Industry	Depreciation per Employee (EUR/FTE)
Electricity distribution firms	47,633
Other industries	7,840

2.2.2 Monopoly and Efficiency

As it would be uneconomic to have more than one provider, by definition this single provider will be the (natural) monopolist. The presence of a monopolist, however, gives rise to market power and in turn leads to market failure.[2] Economic theory predicts that the presence of a monopolist leads to inefficiencies in the allocative and the productive sense as well as to a suboptimal supply of quality. This section discusses the two efficiency problems. Then, in the next section, attention is given to the problems occurring in the area of quality.

Allocative efficiency is measured by the total social surplus, which is the sum of producer and consumer surplus. Producer surplus is given simply by profits i.e. the difference between revenues and costs. Consumer surplus is provided by the difference between consumer willingness to pay and the going price at the given output. As long as the willingness to pay by consumers is higher than the marginal costs of producing the unit, an increase in output increases total surplus and consequently allocative efficiency. Maximum allocative (or Pareto) efficiency (given the existing demand and supply conditions) is reached when costs and willingness to pay are equal at the margin. In the Pareto optimal state there does not exist another feasible state where one party is better off (on a higher utility level) and no other party is made worse off. Then, an optimal allocation of scarce resources is reached and society as a whole will be best off.

However, a monopoly firm will generally not supply an output that corresponds to the Pareto efficient level. In line with Spence (1975), this can be demonstrated as follows. Assume that profit π is given by:

$$\pi = x \cdot P(x) - c(x) \tag{2-1}$$

Here, *x* is the output generated, *P (x)* is the inverse demand, and c*(x)* stands for costs. The monopolist will aim to maximise his profits; this requires that marginal profits are zero:

$$\begin{aligned} &\pi_x = 0 \\ &\Rightarrow c_x = P + x \cdot P_x \end{aligned} \tag{2-2}$$

As can be seen, profits are maximised if marginal revenue equates marginal costs. For social welfare to be maximal however, optimum output requires that prices equate marginal costs. Let total surplus be denoted by:

$$W = S + \pi = \int_0^x P(v)dv - c(x) \tag{2-3}$$

Pareto optimality would be achieved if total surplus is maximised, which requires that:

$$\begin{aligned} &W_x = 0 \\ &\Rightarrow P = c_x \end{aligned} \tag{2-4}$$

As can be observed, the monopoly and social optimum will generally not coincide and as a result, there will be allocative inefficiency. This problem gives rise to the need for intervention through regulation. Regulation, in this context, can be defined as taking actions (in principle by government) in order to achieve efficient outcomes i.e. outcomes similar as those that would be attained if competition had been possible.

With respect to the problem of allocative inefficiency, an action the regulator could take is to simply force the monopoly firm to supply at a price that is equal to marginal costs. However, this apparently simple solution suffers from two major problems. Firstly, it is unlikely that the regulator will know what the marginal costs really are. The regulator faces the problem of information asymmetry between itself (the principal) and the regulated firm (the agent). Assume for the moment that the regulator has perfect information and is capable of setting prices equal to marginal costs. Then, there is still a second problem as pricing at marginal cost would lead the firm to running a loss. This can be demonstrated as follows. Let costs be equal to the sum of fixed costs (*FC*) and variable costs (*VC)* where only the latter varies with output *x*. Total costs (*TC*) and average costs (*AC*) are given by:

$$TC = FC + VC \tag{2-5}$$

$$AC = TC/x \tag{2-6}$$

Pricing at marginal costs $p=TC_x$ implies profits equal to:

$$\pi = x \cdot TC_x - TC \tag{2-7}$$

And profits per unit output would then be:

$$\pi/x = TC_x - TC/x < 0 \tag{2-8}$$

As can be seen, profits will always be negative as one is dealing with scale economies: If average costs decrease with output it holds that average costs are higher than marginal costs.[3] Thus, pricing at marginal costs in the natural monopoly case is not feasible as the firm then would not be able to recover its costs. Consequently, the firm will not be financially sustainable and possible bankruptcy may result. There are in principle two options to deal with this problem. One option is to provide the firm with a subsidy equal to the amount of losses that the firm would incur if prices were set at marginal costs. This subsidy could be financed through taxation but this creates economic distortions elsewhere in the economy. Also, the presence of a direct subsidy from the regulator to the firm may give rise to problems of regulatory capture (see also section 2.3.3). The second option, which is generally applied, is to set prices not at the level of marginal costs but at that of average costs. This is the second-best option, which ensures that the firm remains in business, and service to society is maintained.

In addition to allocative inefficiencies, monopoly also creates inefficiencies in the productive sense. Productive efficiency implies that the firm is producing a given output with the least amount of inputs. Low productivity results from excessive amounts of certain inputs or from using the wrong input mix such as suboptimal substitution between labour and capital.[4] Leibenstein (1966) denotes the waste in production and the wrong choice of production techniques as lack of “X-efficiency”. Under competition, productive efficiency will be reached because firms have an incentive to produce at lowest possible costs as this leads to higher profits. Under monopoly, however, there is no competitive pressure and this results in slack in costs control and effort by the firm’s managers and personnel. Due to this, the firm will operate at a too high cost level, i.e. it could well decrease costs (inputs) at the given level of outputs. When the factor time is taken into account, another form of inefficiency arises, so-called dynamic inefficiency: Over time, there will be no or insufficient improvement in products and production techniques.

In summary, the second-best solution for the allocative efficiency problem is to set prices equal to average costs. However, this does not solve the issue of productive efficiency. If the underlying costs are inefficiently high, this will be directly reflected in higher prices. One of the main problems in regulation is how to design and implement appropriate incentives to promote productive efficiency. This issue will be studied extensively further in this thesis.

2.2.3 Monopoly and Quality

The problem of monopoly is not only related to allocative and productive (in)efficiency, but also to quality. In general, the monopolist’s choice of quality does not coincide with the social optimum. The economic literature on quality under monopoly and regulation is limited when compared to that on price regulation. Generally, two streams of literature deal with quality. The first stream studies quality outcomes under monopoly situations and the problems involved in correcting unfavourable outcomes through regulation. The second stream of literature deals with quality in a non-monopolistic setting (competition, duopoly, oligopoly) and considers issues like market imperfections created by informational asymmetries about quality, quality innovations, the role of advertising and R&D, etc.[5] Given the scope of this thesis, the second literature stream is excluded from this research.

One of the earliest and most fundamental questions assessed in the quality regulation literature is: Does market structure influence quality outcome? The difference in price outcomes between monopoly and competition was well known. It was, however, not clear whether comparable differences would occur with respect to quality. Early articles by Swan (1970) and Lancaster (1975) studied the level of product variation by a monopolist and find that – similar as under perfect competition - an optimal range of varieties will be produced. However, White (1977) showed that this result is driven by the (too strict) assumptions made in their analysis: The Swan-Lancaster results only hold true if the monopolist is able to almost perfectly price-discriminate, i.e. to force some buyers to pay higher prices for the same good than other buyers do. If the monopolist cannot perfectly discriminate – which is generally the case – he will in general not offer an optimal range of varieties to everyone, but instead will produce non-optimal varieties for some consumers or may even refuse to provide any satisfactory varieties to some consumer groups. An important consequence of this is that the adverse welfare effects of monopoly are not only limited to allocative and productive inefficiencies, but that exercise of monopoly power also causes welfare losses at the quality front.

Having determined that monopoly leads to suboptimal quality levels, the next question is how to counteract such undesired outcomes. This issue is examined first in the seminal papers by Nobel Prize Laureate Michael Spence (1975) and Eytan Sheshinski (1976) who, independently from each other, analyse price, quantity, and quality under monopoly and the impact of regulation.[6] In their models, they assume that a single product is produced and quality is uniformly supplied to all consumers. They find that the monopolist will provide either higher or lower quality than the optimal level, depending on demand conditions. If consumers' willingness to pay for quality decreases (increases) at higher demand levels, then quality will be lower (higher) than the social optimum. The source of this failure is, as Sheshinski (1976) puts it:

> *"When the monopolist extracts from each consumer his value of an upgraded quality, the equilibrium quality level is socially optimum, but when all consumers are equally assessed by the marginal consumer's evaluation, decisions are distorted."*

The monopolist provider decides on the basis of the demand of the marginal consumer whilst for social welfare, it is the average consumer that matters. As the marginal consumer is not likely to represent the average consumer, the monopoly choice of quality will generally not coincide with the social optimum (if known at all). This is an important result and deserves further explanation. In doing so, the model by Spence (1975) is useful to consider. Assume that the monopolist firm produces x units of a product of quality q and price p. The costs to produce quality are c and inverse demand is given by P. Consumer surplus, revenues, profits, and total surplus are denoted as respectively CP, Rev, π and W, and defined as follows:

$$CP = \int_0^x P(v,q)dv - x \cdot P(x,q) \tag{2-9}$$

$$Rev = x \cdot P(x,q) = p \cdot D(p,q) \tag{2-10}$$

$$\pi = Rev - c \tag{2-11}$$

$$W = CP + \pi \tag{2-12}$$

For a given quality level, welfare is maximised with respect to quantity if $P=c_x$. But monopoly profits are maximised if $P - c_x = - x \cdot P_x$ which is the familiar outcome as the firm exploits its monopoly position. If quality is a decision variable, then for a given quantity x, welfare is maximised when:

$$\frac{\partial W}{\partial q} = \int_0^x P_q dv - c_q = 0 \tag{2-13}$$

But the firm's profits are maximised if:

$$\frac{\partial \pi}{\partial q} = x \cdot P_q - c_q = 0 \tag{2-14}$$

Thus, the firm's private optimum and the social optimum would only coincide if:

$$\int_0^x P_q dv = x \cdot P_q \tag{2-15}$$

Alternatively, this can be written as:

$$\frac{1}{x}\int_0^x P_q dv = P_q \tag{2-16}$$

The above equation implies that for quality levels to be socially optimal, the average and marginal consumer valuation of quality should be equal. In reality, however, the firm's quality choice is driven by the profit maximising motive – this implies that quality is set on the basis of the marginal consumer instead of the average consumer. The firm's private quality choice would thus only be optimal if the marginal consumer would represent the average consumer. But as this is not likely to apply, the monopoly firm's choice of quality levels will generally not coincide with the social optimum. Simply stated, the monopolist is interested in how quality changes the marginal willingness to pay by consumers whilst instead the average change in willingness to pay should have been considered. This leads the monopoly equilibrium quality level to differ from the socially optimal one.

The next question that Spence and Sheshinski investigate is whether this monopoly quality will be higher or lower than the optimal one. The monopolist will maximise profits and hence choose a quality level in such a way that the derivate of profits with respect to quality is zero. The issue then is how a change in quality affects social welfare. If the change in welfare in response to a quality increase is positive then quality is too low. In practical terms, this means that the firm stops investing in quality too soon. Conversely, when quality increase leads to lower welfare then the monopoly quality is too high.

The full analysis will not be replicated here, but only a summary of the main conclusions obtained by Spence and Sheshinski is presented following Varian (1993).[7] Firstly, one would need to consider the relative impact of a quality change on the marginal versus the average consumer. If the average consumer values the change in quality more than the marginal consumer then the monopolist under-provides quality. Secondly, the change in consumer surplus as a response to the change in the demand curve resulting from quality deviations has to be taken into account. In Varian's terminology, one should look at the effects of both the "shift" and "tilt" in the demand curve. The shift does not change consumers' surplus at all, only the tilt matters. Changes in quality that only shift the demand curve have no effect on welfare as they do not affect the difference between average and marginal valuation. To affect the average and marginal consumer differently, the change in quality must affect the slope of the demand curve; only variables that tilt the demand curve will increase or decrease welfare. Depending on the changes in consumer quality preferences, this leads to an increase and decrease in consumer surplus respectively. In practice, collecting data in order to empirically assess the impact of quality changes on the demand curve is quite difficult. In fact, Spence (1975) considers this one of the three main information problems that complicates quality regulation. Later, in chapter four, this problem is, among others, explored in more detail.

2.3 Regulation of Monopoly

2.3.1 The Regulatory Problem

Regulation can be considered as a solution for the natural monopoly problem. It can be defined as a means to counteract the adverse monopoly effects by creating mechanisms driving towards monopoly outcomes that maximise total social welfare. This objective can be divided into two components: Achieving maximum economic efficiency and attaining an optimal quality level. At the same time, however, the regulator also needs to take into account the interests of the (monopoly) firm. Investors require a sufficiently high return on their investment. Too low price levels reduce profitability and may make it unattractive for investors to enter the business. The regulator therefore needs to set prices at sufficiently high levels in order for the investor to earn a reasonable rate-of-return. This is also known as the participation constraint.

One way to look at the financial sustainability objective is to observe the amount of risk to which the firm is exposed. If the firm is exposed to higher risks, its costs of capital will increase and consequently, it will require a higher return on investment in order to remain viable. This relation between risks and returns is well known in finance theory.[8] Alexander and Irwin (1996) observe that a move from rate-of-return regulation to price-cap regulation leads to an increase in risk to the firm. Under rate-of-return regulation, if a firm's costs rise, the firm would seek for a price review and typically be granted one within a year or so, so its profits would not change much. Under price-cap regulation on the other hand, price levels follow a predefined path and now, the firm's profits will fall because it cannot raise its prices to compensate for the cost increase (at least, not until the next regulatory period). This tends to raise the firm's costs of capital for which Alexander and Irwin (1996) provide some empirical evidence. In their study, they measure the impact of the type of regulation on the firm's capital costs, measured by a statistic called the firm's beta.[9] Their survey shows that firms that are regulated on the basis of price-caps tend to have a beta that is higher compared to firms regulated under rate-of-return regulation. This difference in risk perception, according to the authors, is associated with approximately one extra percentage point of capital cost. The survey unfortunately does not differentiate between different types of price-cap regulation or rate-of-return regulation. It is likely that not only the choice for either rate-of-return regulation or price-cap regulation will affect the firm's costs of capital, but that other specifics of the regulatory approach will also influence the firm's risks and consequently its costs of capital.

The sustainability objective requires that the firm remains in business i.e. does not go bankrupt. In a competitive industry, firms that are run badly compared to their competitors become unprofitable and will exit the market; their places will be taken over quickly by other firms (competitors or new entrants). This logic, however, does not apply directly to regulated monopoly firms. There are primarily two reasons for this. Firstly, one should take into consideration that the firm's price is set by the regulator based on assumptions made under information constraints. Erroneous regulatory estimates of the firm's efficiency, capital costs, demand forecasts, etc. can lead to a wrong price level and therewith adversely affect the firm's profitability. It will be very difficult to justify bankruptcy when this can be (partially) traced back to regulatory misjudgements.[10] The second reason why bankruptcy is highly undesirable is that it will lead to substantial social and political turmoil. An electricity distribution firm is different from most firms in the fact that it plays a vital role in providing an essential service, the service of electricity supply, to society and the economy. Therefore, in case of pending bankruptcy, government is likely to intervene by changing the regulatory policy or by providing financial funds. This, for example, happened during the California Power Crisis (De Vries 2004, p. 44-45).

In summary, economic sustainability implies that the regulated firm should be able to finance its operations and make any required investments, so that it can continue to operate in the future. Shareholders should receive adequate returns on their investments while investors require enough security to put their money in assets that, once bought, become sunk and need to produce secure revenues over a large number of years. At the same time, consumers may not be overcharged – they are entitled to protection as captive users. Service affordability as well as quality should thus be ensured at socially optimal levels.

2.3.2 Franchise Bidding

Generally, public policy makers have adopted regulation as the primary mechanism to counteract monopoly price and quality outcomes while at the same time ensuring continuity of the firm. Regulation, however, has also been criticised by some for being the wrong choice to solve the "natural monopoly problem". In particular, in his influential paper, Demsetz (1968) criticises the theory of natural monopoly as follows:

> *"…for it [regulation] fails to reveal the logical steps that carry it from scale economies in production to monopoly price in the market place."*

And further:

> *"…scale economies in servicing demand in no way imply that there will be one bidder only. There can be many bidders and the bid that wins will be the lowest. The existence of scale economies in the production of the services is irrelevant to a determination of the number of rival bidders."*

Demsetz argues that even if technology necessitates a single provider, the process by which this monopoly firm is selected can make sure that the market outcome is not one of a monopoly. He claims that by letting multiple firms compete for the market and awarding the contract to serve to the lowest bidding firm for a long enough period, the social optimum can be reached. Effectively, Demsetz argues that rather than competition *in* the market, one should apply competition *for* the market and this in turn eliminates the need for regulation. The bidding process for the franchise, which has come to be known as the Demsetz Auction, is proposed as a superior solution for dealing with the natural monopoly phenomenon.

The Demsetz Auction as a replacement for regulation has in turn been heavily criticised. In particular by Williamson (1976) who studied the effectiveness of franchise bidding in the provision of public utility services for Cable Television. He finds that:

> *"…in circumstances where franchise bidding predictably and actually converges toward regulation, the purported advantages of franchise bidding as compared with regulation are problematical."*

Williamson's main critique is aimed at the problems of durability and uncertainty that Demsetz in his analysis considered to be "irrelevant complications". According to Williamson, over time, costs and demand conditions will change and the price of today might not be optimal in the future. These uncertainties cannot be captured completely in contracts through clauses or provisions.[11] Furthermore, there are transaction costs associated with implementing the contract and monitoring the conditions. Thus, the costs of the Demsetz Auction are not limited to a once-for-all event but will continue to persist and even grow over time. In particular, problems occur in subsequent auctions when assets are retransformed from existing franchisees to successor firms. There is an enormous problem in the (re)valuation of these assets. New bidders may not be in an equal position compared to existing franchisees who have much more knowledge as they have been operating the assets for a long time already. This leads Williamson to conclude that the problems faced under franchise bidding can even be more severe as those experienced under regulation. Although the difficulties of franchise bidding can be mitigated, this will come at the costs of an extensive regulatory/arbitration apparatus. According to Williamson (1976), franchise bidding may well function better than regulation in some cases e.g. airlines and postal delivery. However, it is definitely not the most appropriate solution in all circumstances, as Demsetz is believed to argue.

2.3.3 Capture Theory

Notwithstanding the critique on the effectiveness of the Demsetz Auction, there are other and perhaps more fundamental critiques on regulation. In his same paper, Demsetz (1968) claims that the monopolist firms prefer regulation because it provides them with a comfortable protection against the market. This, as Demsetz argues, is exactly the reason why firms actively promoted regulation as the proper solution for the natural monopoly problem. This reasoning is in line with that of the so-called Chicago School which takes the view that market power does not exist and if it does, it results precisely from government intervention: It's not that market power prompts government intervention, but the exact opposite – government intervention creates market power, protecting the interests of firms and not those of consumers (Cabral 2000, p. 10). Proponents of the Chicago School believe that there is no such thing as a benevolent regulator i.e. one acting in the interest of society alone. For example, George Stigler, one of the main advocates of the Chicago School and Nobel Prize Laureate adopts the following view of regulation in his classic paper on regulation (Stigler 1971):

> *"A central thesis in this paper is that, as a rule, regulation is acquired by the [regulated] industry and is designed and operated primarily for its benefit."*

Subsequent authors including Posner (1971), Posner (1974) and Peltzman (1976) consider this problem of "regulatory capture" where the regulator's objective is other than maximising social welfare. Peltzman (1989) summarises the characteristics of the capture theory as follows:

- Compact, well-organised groups (usually producers) will tend to benefit more from regulation than broad, diffuse groups (usually consumers);
- Regulatory policy will seek to preserve a politically optimal distribution of rents across this coalition. Thus, over time, the policy will tend to offset changes in this optimal distribution arising from shifts in demand or costs conditions;
- Because the political payoff to regulation arises from distributing wealth, the regulatory process is sensitive to deadweight losses. Policies that reduce total wealth will be avoided because they reduce the political payoff from regulation.

Capture theory provides a rather pessimistic view of regulation that as one may assume – perhaps naively – was established in the first place to the benefit and protection of consumers. This view is perhaps driven by the recognition of human weakness. Regulators, after all, are also human beings who can misuse the trust that is put in them. An example is the problem of "revolving doors" where regulators tend to favour firms as they anticipate future management positions in these firms. Trust can also be abused more bluntly such as when bribes are accepted. In particular, when the regulator is vaguely mandated and has much discretion, there is a risk that he may misuse public trust at his personal advantage. This is also the reason why regulators are usually not permitted to make lump-sum transfers (subsidies) to the firm (Laffont and Tirole 1993). When there is higher risk of regulatory capture, it is desirable to limit the discretion of the regulator, in particular the mandate of lump-sum transfers to the firm. This can reduce the risk of undesired flow of funds from the public to firms (and then part of it back to the regulator). The regulator could also indirectly achieve such a lump-sum transfer, by allowing an unjustified increase in prices. However, as Armstrong et al. (1994) note, this is less practical as consumers are generally more aware of high prices as opposed to transfers from the general taxation revenue.

In summary, two main lines of critique on regulation can be identified. Firstly, the Demsetz thesis that states that there are other and more effective means to deal with the economic problem of the natural monopoly than regulation. Given the scope of this thesis – which deals with the problem of setting up an effective regulatory mechanism rather than inventing surrogates for it – the assumption is that regulation is the best option one has. Although the problems attached to regulation should be acknowledged, the predominant opinion here is that these problems are of a less severe nature than those under a system of Demsetz Auctions. In fact, as far as known, there is no instance where discontent with regulation has led it to be replaced by a Demsetz Auction – at least, not in the case of electricity distribution.

The second line of critique is that regulation may lead to capture. This problem will be ignored further in this thesis and instead, the view will be adopted of the benevolent regulator who acts purely in the interest of society. There are two motivations for this. Firstly, there is no indication that regulatory capture is a problem within the scope of this research. This thesis considers cases where one moves towards stricter regulatory regimes with a stronger focus on higher efficiency performance. Here, the relationship between the regulator and the industry is often found to be hostile – in sharp contrast with the dreaded situation of regulatory capture. Secondly, usually there are sufficient checks and balances set in place to make sure that regulatory capture is prevented. Relatively simple measures can already mitigate the risk of regulatory capture. These measures include making the regulatory processes more transparent, prohibiting conflicts of interest, providing effective arrangements for appeal and public scrutiny of the regulator's budget, conduct, and efficiency (Smith 1997). Thus, the assumption is that sufficient guarantees are built in to further ignore the capture problem and to adopt a more amiable view of regulation: That of the benevolent regulator acting to protect consumers from abuse by firms with substantial market power, to support investment by protecting investors from arbitrary action by government, and to promote economic efficiency and quality.

2.4 Rate-of-Return Regulation

Traditionally, regulators have tended to set prices on the basis of average costs as observed in the past or as projected by the regulated firm. This is known as rate-of-return regulation. Here, prices are set in such a way that the firm can recover all its (projected) costs including a "fair" return on investment.[12] Rate-of-return regulation has a long tradition in the US. Already in 1898, the Court stated in the well-known Smyth versus Ames case:

> *"The company is entitled to ask for a fair return upon the value of that which it employs for the public convenience."*

At the same time, the Court ruled that consumers also deserved protection:

> *"…while the public is entitled to demand…that no more be extracted from it…than the services are reasonably worth."*

In the historically important Federal Power Commission versus Hope Natural Gas Co. case of 1944, the Court ruled that regulated firms are entitled to a "just and reasonable" rate-of-return, so that the firm is able:

> *"…to operate successfully, to maintain its financial integrity, to attract capital, and to compensate its investors for the risks assumed."*

The Hope ruling became one of the main foundations for the subsequent decades of American style of regulation and it has been applied to various industries such as electricity, water, gas, telecom, etc. In Europe, there was a different development in the regulation of public utilities. Here, public ownership was the predominant form of regulation. Regulation however, in essence, was similar to US-style rate-of-return regulation (Newbery 1997).

Under rate-of-return regulation, the revenue requirement of the firm is based on the firm's accounting costs during a test year (Liston 1993). These accounting costs include operating costs, taxes, allowances for depreciation, and allowed returns. The allowed return is a "fair" or “reasonable” rate-of-return – based on an estimate of the firm’s costs of capital – multiplied by the rate base that consists of the un-depreciated portion of investments relevant to regulated operations. Based on this allowed revenue, the regulator determines a tariff structure to recover the aggregate costs. Tariffs remain the same for a period until they are reviewed in the next round. Such a new round may be induced by the firm, if it finds that its costs have increased and would like to see prices reflect these changes, or by the regulatory commission or other parties (e.g. consumer representation groups) if these feel that the firm is earning too high returns (Whitaker 2002).

Rate-of-return regulation effectively sets price equal to average costs. As has been noted earlier, marginal cost pricing is not feasible, as this will lead to a financial loss; setting the price equal to average costs is the second-best option. Rate-of-return regulation partially solves the allocative efficiency problem but at the same does little for productive efficiency. This has been one of the main points of critique on rate-of-return regulation during the last four decades. In addition, rate-of-return regulation has also been criticised for the high administrative costs that go with it. These two streams of critique are now reviewed.

2.4.1 Limited Efficiency Incentives

One of the earliest and most influential publications criticising the weak efficiency properties of rate-of-return regulation is by Harvey Averch and Leland Johnson (1962) [Averch-Johnson]. In their analysis, they point out the adverse effects of setting the fair level of return on the basis of capital investment. Their point can be demonstrated as follows. Let the rate-of-return be defined as revenues minus costs for non-capital inputs, divided by the level of investment (*K*). If there is only one non-capital input, labour, denoted as *L*, the rate-of-return is given by $(p \cdot x - w \cdot L)/K$. Here, p and x denote price and output, and w is the price of labour. The firm is free to choose its price, output levels and inputs as long as its return rate does not exceed the fair rate ror_f set by the regulator. This is known as the rate-of-return constraint:

$$ror_f \geq \frac{p \cdot x - w \cdot L}{K} \tag{2-17}$$

Let *ror* be the costs of capital. This can be written in terms of economic profit:

$$\pi = p \cdot x - w \cdot L - ror \cdot K \tag{2-18}$$

The above constraint can be re-written as:

$$ror_f \geq \frac{p \cdot x - w \cdot L}{K} - ror \tag{2-19}$$

$$ror_f - ror \geq \frac{\pi}{K} \tag{2-20}$$

and subsequently it follows that:

$$\pi \leq \left(ror_f - ror\right) \cdot K \tag{2-21}$$

As can be observed, the maximum economic profit of the firm increases when the level of investment K is increased. Without going into the technical details of the Averch-Johnson analysis, their main findings can be summarised as follows.[13] Averch-Johnson show that when the regulator sets the fair rate-of-return above the costs of capital, the following inefficiencies arise:

- The regulated firm uses more capital then if it were unregulated.
- The capital/labour ratio of the regulated firm is inefficiently high for its level of output. That is, the output that the regulated firm produces could be more cheaply produced with less capital and more labour than the regulated firm chooses.
- It is possible that the firm produces less output and charge a higher price than if it were not regulated.
- The regulated firm necessarily operates in the elastic portion of demand, where marginal revenue is positive. That is, the regulated firm never increases its output beyond the point at which marginal revenue is zero.
- The regulated firm produces as much output as possible given its choice of inputs (capital and labour).

The overall conclusion from the Averch-Johnson analysis is that when the rate-of-return is higher than the firm's costs of capital, the firm chooses an inefficient mix of inputs and wastes capital. When the fair rate is set equal to the costs of capital, the firm becomes indifferent among many levels of output and many input combinations, including the option of closing down. If the rate-of-return is set lower than the costs of capital, closing down is the only alternative. The consequence of this is that the regulator will always need to set the rate-of-return higher than the costs of capital in order to maintain the firm's financial sustainability. Thus, rate-of-return regulation in principle always leads to inefficiency in production: The firm chooses inefficiently high levels of capital in order to boost its profits. This is known as the Averch-Johnson or overcapitalisation effect.

These results obtained by Averch-Johnson rank among the most important ones in regulation theory. Crew and Kleindorfer (2002) portray the Averch-Johnson paper as follows:

> *"Although many authors have sought to discredit this paper it is one of the most highly cited and influential papers in regulatory economics."*

Similarly, Braeutigam and Panzar (1993) note that:

> *"Although a number of analysts have questioned the validity of the Averch and Johnson model and some have attempted to modify it to better reflect reality, many of the basic concerns about rate-of-return regulation today are remarkably similar to those suggested by Averch and Johnson more than 30 years ago."*

Apart from the overcapitalisation problem, rate-of-return regulation is also generally criticised for its lack of incentives to make efficient use of inputs i.e. use the least amounts of inputs to produce the given outputs. As noted earlier, this is part of Leibenstein's X-efficiency problem where lack of competition leads to slack in costs control and effort by the firm's managers and personnel. The reason for this is that setting the price equal to average costs effectively guarantees a predetermined profit, irrespective whether the firm is operating efficiently or not. Baumol and Klevorick (1970) were amongst the first to analyse this problem. They argued that as under rate-of-return regulation all profits in excess of the predetermined return are ruled out, this:

> *"…precludes extraordinary rewards for extraordinary entrepreneurial accomplishment."*

Simply stated, firms do not have any reason to become (more) efficient as doing so does not generate any rewards (in the form of higher profits) to them anyhow. If the firm reduces its costs, it earns higher profits but these profits will be taken away in the next round of tariff setting. To deal with this problem, Baumol and Klevorick propose to keep prices fixed between two price reviews. By formalising the period between these reviews i.e. introducing a regulatory time-lag, firms retain the profits resulting from costs reductions and thus strong incentives for efficiency improvement are created. Excess profits can be earned as through research and innovation the firm can reduce its costs below the level prevailing when prices were last set by the regulator. Consumers also profit from these efficiency improvements. At the next regulatory review, prices will be adjusted downwards to take into account the firm's improved technology and the process begins all over again.[14] The length of the regulatory lag is the trade-off between higher efficiency when the lag is longer, and more surplus transfer to consumers when the lag is shorter.

The Averch-Johnson findings, inefficient choice in capital inputs, and the Baumol-Klevorick analysis, the absence of incentives to operate efficiently, point out the general drawbacks of rate-of-return regulation: A lack of incentives for productive efficiency i.e. for selecting the optimal level and combination of inputs. On the one side, setting a fair return on the basis of invested capital creates perverse incentives to overcapitalise. On the other side, setting prices equal to costs blunts the firm's incentive to produce at minimum costs. In addition to other problems, which will be described further, both observations played a key role in the general recognition of the weak efficiency properties of rate-of-return regulation and the subsequent move towards regulation systems that provide stronger efficiency incentives. Indeed, Baumol and Klevorick's proposal to formalise the lag between price reviews as a means to promote productivity is one of the key concepts of price-cap regulation as it is widely applied today.[15]

2.4.2 Administrative Burden

In the US, rate-of-return regulation has not only been criticised for its weak efficiency properties but also for the high administrative burden that it engenders. The main feature of rate-of-return regulation is that prices follow costs. Usually the regulator has discretion in determining which cost elements should form part of these costs and which should not. As regulators are often under public and political pressure to keep prices low, they tend to be highly critical in assessing which part of the firm's costs should be allowed to enter the tariffs. This can lead to highly intrusive forms of regulation – micro-management – where the regulator indirectly dictates the firm's policies. In determining whether or not, and which portion of costs should be allowed to enter the tariffs, regulators often have made use of the so-called "used-and-useful" principle. Huygen (1999, p. 72-73) presents some examples of this principle: In one case, the utility commission in Connecticut considered the salary level of an individual employee too high. In another case, the Utah Power & Light Company was disallowed to pass on the cleaning costs resulting from a leak of toxic gasses through the tariffs but was allowed the costs of a management audit.

Although there have been substantial debates about which costs should enter the tariffs, the main controversies according to Huygen (1999, p. 73) emerged at the capital costs side, in the determination of the allowed rate base and the "fair" level of return. This led to substantial debates – usually in the form of time-consuming hearings cases – and in turn to a large administrative apparatus dedicated to the implementation of rate-of-return regulation. In the US, Public Utility Commissions are, among others, responsible for determining tariffs for the utilities under their jurisdiction. These Commissions often employ a large number of staff, which continued to grow from 2,599 in 1936 to 81,000 in 1977 (Breyer 1982).

According to Whitaker (2002), the costs of rate-of-return regulation were further increased as a result of the oil crises and high inflation. These developments led firms to more frequently appeal for a tariff increase resulting in a dramatic increase in rate cases in the 1970s and 1980s. At the same time, consumers were increasingly dissatisfied with the higher prices and got more involved in the rate cases, resulting in even more judicial and lengthy procedures and ultimately higher regulatory costs. This eventually led to strong dissatisfaction with the rate-of-return regulation system. When liberalisation was introduced, this was taken as the opportunity to get rid of the time and money consuming system of rate-of-return regulation and adopt price-cap regulation, which was believed to be a more efficient regulatory system.

In Europe, there has not been a long and formal regulatory tradition as in the US. Here, the move towards price-cap regulation can be considered a by-product of the general power sector reform, which is characterised by the introduction of competition, change of ownership, and establishment of new regulatory institutions to organise these reforms and deal with the remaining monopoly networks (Jeunemaitre and Matheu 2001). European legislators were aware of the necessity to explicitly regulate the remaining network monopolies (as well as other activities that remained non-competitive at least in the shorter term) but at the same time, they did not want to introduce an enormous and costly regulatory apparatus as observed in the US. There was thus a deliberate choice for "light-handed" price-cap regulation as opposed to the "heavy-handed" and as too costly perceived rate-of-return regulation of the US.

2.4.3 Rate-of-Return Regulation and Quality

Much of today's increased attention for quality regulation is driven by the anticipation of quality problems resulting from the change towards price-cap regulation. The question then arises whether these concerns have not been there under rate-of-return regulation and if not, why this has been the case. A passage from the influential book by Alfred Kahn (1970) helps to explore this question:

> *"But it is far more true of quality of service than of price that the primary responsibility remains with the supplying company instead of with the regulatory agency, and that the agencies, in turn, have devoted much more attention to the latter than to the former. The reasons for this are fairly clear. Service standards are often much more difficult to specify by the promulgation of rules."*

Kahn's observation suggests that the perceived difficulties of regulating quality have been the reason for regulators to "leave quality to the firms". Regulators seem to have recognised the importance of quality, but at the same time did not see an explicit need for quality regulation. But why were regulators so comfortable in ignoring quality issues? The answer perhaps lies in the natural tendency to oversupply quality under rate-of-return regulation. As was seen in paragraph 2.4.1, there is a tendency towards overcapitalisation if the fair rate-of-return is set higher than the firm's costs of capital.[16] If quality is capital intensive, then quality levels will automatically tend to be high and there will be less need for explicit quality regulation. Then, rate-of-return regulation can be considered as a substitute for quality regulation (Spence 1975).[17]

However, overcapitalisation does not necessarily mean that quality is at an optimal level. From a theoretical point of view, quality would only be optimal if the costs to produce it and consumer demands for it would be equal at the margin.[18] In the rate-of-return regulation case, one may expect an oversupply of quality (so-called gold plating) i.e. a quality level higher than the optimum. This would come at additional costs and thus a higher price – consumers will be paying a too high price for a too high quality level. Another observation by Alfred Kahn, more than thirty years after the one he made in 1970, is very relevant in this respect (Kahn 2002):

> *"But there is reason to believe that it [high reliability] has come at too high a price. There is substantial evidence that the [high quality] standards were selected by engineers to make their lives easy rather than to save customers money."*

The concern about gold plating is supported by different empirical studies including those by Shipley et al. (1972), Telson (1975) and Bental and Ravid (1982). These authors use US data to compare the costs of supplying higher system reliability against the benefits of doing so and find that – at given existing reliability levels – the former is substantially higher than the latter. The main reason for this is the choice of too high reliability standards. For example, Telson (1975) calculates that reliability in New York State should be reduced by a factor five to arrive at optimal levels. More recently, a study by the Dutch energy regulator showed that the total costs of interruptions caused annually to consumers in the Netherlands are only a small fraction (2.5 percent) of the total system costs (DTE, 2004c). Although this cannot be considered conclusive evidence, it does point out into the direction of too high rather than too low reliability levels. Studies like these seem to confirm the view that rate-of-return regulation not only resulted in low productivity but also led to inefficiently high quality levels in the electricity industry.

2.5 Price-Cap Regulation

2.5.1 Incentive Regulation

The previous section discussed rate-of-return regulation and some of its disadvantages. The discussion now moves to some of the alternative price control mechanisms which have been proposed to replace rate-of-return regulation and which can be grouped under the term incentive regulation.[19] The main difference from rate-of-return regulation is the strong emphasis on the use of incentives to induce the regulated firm to operate in a certain desirable way. This is done by making the firm (partially) claimant of the residual gains resulting from better performance. That is, if the firm operates in a manner that is more consistent with the regulatory objectives, it is allowed a part of the benefits accruing from this better performance. Laffont and Tirole (1993) describe incentive regulation as the firm and regulator engaging into a contract. This contract (the incentive scheme) takes the form of the regulator paying the firm's costs (*c*) and then paying a net transfer *tr* to the firm:

$$tr = \alpha - \beta \cdot c \qquad (2\text{-}22)$$

Here, α is a "fixed fee" and β is the fraction of costs born by the firm. Alternatively, one could adopt the convention that the firm pays for its costs and that the government reimburses a fraction of $1 - \beta$ of the costs and gives a fee of α. The power of the incentive scheme is given by β. Two polar cases of regulatory contracts exist: At one extreme, pure rate-of-return regulation contracts (β=0) where all the firm's costs are fully reimbursed, this is an extremely low-powered incentive scheme. At the other extreme, fixed-price contracts (β=1) completely remove the link between the firm's costs and allowed income; this pure price-cap regulation scheme is an extremely high-powered incentive scheme. The firm is the full residual claimant i.e. it is allowed to keep all costs savings. In practice, incentive schemes will lie anywhere between these two polar cases.

Initially, proposals for incentive schemes dealt with the problem of how to make the firm operate more efficiently. Loeb and Magat (1979) proposed a scheme under which the regulated firm is allowed to freely set the price and is allowed to keep all the revenue in addition to a subsidy equal to the value of the consumer surplus. Because the firm will maximise profits, it will set a price at which the sum of its profits plus the subsidy equal to the consumer surplus is maximised. But this is exactly the price at which producer and consumer surplus is largest and is by definition the optimal outcome. The firm will also produce efficiently because reductions in costs translate into higher profits, which it is allowed to keep. The Loeb-Magat scheme is incentive compatible, i.e. it assures that the firm's profit maximising motive leads to an outcome that it desirable from the regulator's point of view. Incentive compatibility is achieved by internalising into the firm's decision making process the effects of its choices on consumer surplus. The obvious problem, however, is that although the scheme (theoretically) leads to an optimal outcome in terms of efficiency, the distributional properties of this outcome are at least debatable. Under the Loeb-Magat scheme all surplus is attributed to the firm while consumers are left with zero surplus. Clearly, this outcome will not be sustainable in practice due to distributional concerns.

An important assumption of the Loeb-Magat scheme is that the regulator knows consumer demand. Finsinger and Vogelsang (1985) relax this assumption. In their scheme, the firm earns an approximation of the change in total surplus between two previous periods. This approximation is based on the prices and quantities that the regulator observes in each period. After a series of periods, the firm converges towards the optimal outcome. A problem here is that as over time conditions change, the optimal outcome may also change and consequently it will never be reached. Sappington and Sibley (1988) extend the Loeb-Magat scheme dynamically by considering multiple periods under which the regulator uses information from the previous period to set the level of the subsidy for the next period. They call this the incremental surplus subsidy scheme. Here, the firm is only awarded the increase in surplus between the latest two consecutive periods. This also leads to an optimal outcome but the level of the subsidy is less than under the Loeb-Magat scheme.

2.5.2 New Regulatory Economics

The Loeb-Magat scheme and its successors assume that the regulator has perfect information. For example, the Loeb-Magat scheme could only be applied when the regulator has information about consumer demand in order to calculate the consumer surplus. In practice, this is not a realistic assumption and the regulator will not be able to optimally design the incentive scheme. These and other types of informational problems play the main role in what is known as the "new regulatory economics" (Laffont and Tirole 1993, p. xvii-xix). This new regulatory economics is built around a principal-agent framework. Here, the principal (regulator) assigns a certain task to the agent (firm) but information asymmetries between the principal and agent constrain the principal in implementing its preferred policy. Laffont and Tirole argue that this approach should provide "a more rigorous and realistic approach" than the traditional theory of economic regulation which largely neglected the information asymmetry between the regulator and the firm.

The earliest contribution that can be considered part of the new regulatory economics is by Baron and Myerson (1982). They develop an optimal regulatory policy for the case in which the regulator does not know the costs of the firm. As a mean to solve the distributional problem under the Loeb-Magat scheme (producers receive all social surplus and consumers none) they propose to transfer surplus from producers to consumer by levying a lump-sum tax. However, when the regulator does not know the firm's costs, it runs a risk that the tax is set too high and the firm will decline to supply the good. Since the regulator does not know the firm's costs, the regulator cannot set an optimal price but instead must set the firm's price and subsidy as a function of some costs report from the firm. The regulatory policy must satisfy the constraint that the firm should have an incentive to report truthfully the information desired by the regulator. Because of this constraint, there is a welfare loss resulting from the informational asymmetry. They show that the regulator has to face a trade-off between attaining allocative efficiency and minimising undesired distributional effects. Under the Loeb-Magat scheme, an optimal outcome was achieved (largest social surplus) but all surplus was retained by the firm. If the regulator attaches more importance to consumer surplus, a more balanced distribution of surplus can be reached by setting prices above marginal costs and providing the firm with a smaller subsidy.[20]

Laffont and Tirole (1986) extend the Baron and Myerson model by assuming that the regulator has information about the firm's (accounting) costs but not about the level of effort employed by the firm to lower its costs. In their model, the lump-sum transfer to the firm is used as a means to induce productive efficiency. The lack of information prevents the regulator to do this in an optimal way and as a result, the firm can obtain information rents. They show that the regulator then has a trade-off between allocative efficiency, distributional effects and productive efficiency. The regulator's lack of information leads the firm to under-invest in productive efficiency improvements.

Under the framework of the new regulatory economics, the regulator's task is to develop an incentive scheme which, taking into account the different constraints, leads to optimal regulatory outcomes. Laffont and Tirole (1993) identify two types of informational constraints. Firstly, moral hazard refers to endogenous variables, which are not observable by the regulator. The discretionary actions taken by the firm that affect its costs or quality of its product is denoted as "effort". It is "negative effort" that is of most concern, examples are hiring of unnecessary personnel, indulgence in activities that privilege their career potential over efficiency, purchase of material at high prices, etc. The second type of informational constraint is adverse selection, which arises when the firm has more information than the regulator about exogenous variables (e.g. demand and technology) and thus can exploit this informational asymmetry for its own benefit. The presence of moral hazard and adverse selection create a need for information gathering and monitoring by regulators. This, however, goes further than simple accounting information as these do not reveal most dimensions of moral hazard and adverse selection. It is these residual asymmetries of information left by the accounting process that are the topic of study in the Laffont and Tirole analysis.

In addition to information, other constraints are transactional, administrative, and political. Transactional constraints arise from the fact that contracts are costly to write and likely to be incomplete. Future contingencies must be considered; they must be unambiguously specified and the agreement must be monitored and enforced by a court. It is not likely that the regulator can predict all possible contingencies for the whole duration of the contract. Administrative and political constraints are very diverse. Firstly, the scope of regulation is limited and the regulator cannot intervene in other (interfacing) industries. Secondly, not all instruments are available e.g. a lump sum transfer to the firm is usually not allowed or the regulator may not have the legal power to enforce quality standards. Thirdly, there is a limited time horizon in the regulatory contract with the firms. Fourthly, there are a number of procedural requirements that need to be taken into account such as how information is obtained and regulators deal with firms. Finally, political constraints such as shifting responsibilities or impeachments may affect the regulators decisions.

In their 1993 book, Laffont and Tirole make extensive use of economic modelling to derive optimal incentive mechanisms under, among others, static control of a single firm by a benevolent regulator, product market competition, competition for monopoly position through Demsetz auctioning, multi-period regulation, the politics of regulators and the control of regulators by legislators and institutions, and regulatory institutions. One of the main conclusions from their analysis is that there is a basic trade-off between incentives and rent extraction (Laffont and Tirole 1993, p. 39). Being unable to monitor the firm’s efforts (moral hazard problem) and having less information than the firm about its technology (adverse selection problem), the regulator has to promote cost reduction and extract the firm’s rent. The more efficiency is improved (stronger incentives), the more rents will need to stay with the firm. As the firm improves efficiency, it earns more profits but the incentive to improve efficiency would not be there if it were not to keep these profits in the first place. On the one side, a powerful incentive scheme (e.g. fixed-price) provides strong incentives to increase efforts. However, these savings only precipitate in the firm, whereas consumers do not see any benefit. Under a low-powered incentive scheme, on the other side, there are no incentives for effort improvement but there is full rent extraction. Here, any exogenous variation in costs is received by the government (consumers) and not by the firm.

2.5.3 Evaluation of the New Regulatory Economics

After the original publication by Baron and Myerson, numerous publications have followed analysing the regulatory problem from the principal-agent framework.[21] This framework provides valuable insight into the role of information as a source of monopoly rents. However, the new regulatory economics approach has also been criticised. In particular, Crew and Kleindorfer (2002) argue that:

> *"…its contributions are limited in their applicability and fall short of the expectations."*

They present two main points of critique. Firstly, the heavy reliance on what is generally referred to by new regulatory economists as "common knowledge". Common knowledge is the standard assumption that the regulator can obtain key information about the firm. Such information could include the firm's revenue, cost, profits, investments, technical performance, etc. Based on this common knowledge, the regulator designs the regulatory mechanism. However, as argued by Crew and Kleindorfer, this exactly is the Achilles heel of the new regulatory economics as it is not possible for the regulator to obtain common knowledge:

> *"…without a contested discovery process that always leaves him in a state far short of the level of information assumed in these theories."*

Their critique is that when the regulator does not posses this common knowledge, strategic interactions between the regulator and the firm arise which the new regulatory economics fails to address. Indeed, recent financial scandals like Enron, WorldCom, Parmalat, Ahold, etc. confirm that it is difficult indeed for relative outsiders to obtain genuine information about the firm's costs and performance. Firms may try to present modified or falsified information to the outside world, even with the assistance of supposedly independent auditors.

Crew and Kleindorfer's other point of critique relates to the problem of regulatory commitment. Lack of regulatory commitment occurs if the regulator ex post alters the initial regulatory contract due to a variety of reasons – usual political. To induce efficiency, the regulator has to concede some information rents to the firm (trade-off between incentives and rents). The new regulatory economics assume the regulator to be committed: Ex post appearance of excess profits would not cause the regulator to renege on his commitment to the original incentive scheme. But this, according to Crew and Kleindorfer, is difficult as in practice regulators will not allow monopoly rents to be retained by the firm.

In defence of the new regulatory economics, one should note that even if the (sometimes very theoretical) models that underlie it provide limited practical applicability, their basic ideas did have an important influence on regulatory thinking. Particularly, the importance of the fundamental trade-off between incentives and rent extraction and the importance of regulatory commitment have been generally recognised by regulators. This can perhaps be considered the main contribution of the new regulatory economics. Indeed, the essence of price-cap regulation is that in the intermediate period between the price control reviews, the regulator should not interfere – even if the firm enjoys high rents as a result of increasing efficiencies. The negative impact of clawing back these profits is now well understood. This is also reflected in the establishment of independent regulators who are less sensitive to short-term political influence.

2.5.4 Price-Cap Regulation and Quality

Rate-of-return regulation provides incentives to overuse capital inputs and this, given that quality is capital using, results in high quality. In contrast, price-cap regulation gives firms an incentive to cut costs and this raises the concern that (part of the) cost reductions may be achieved through adverse quality reductions. Theory confirms this concern. Spence (1975) and Sheshinski (1976) already showed that where price is fixed or taken as given, the monopoly firm will always set quality too low. Subsequent publications have studied the quality problem under price-cap regulation in more detail.

Fraser (1994) examines the relationship between price-cap regulation and the reliability of supply of a monopolist. In his analysis, the X-factor in the price-cap represents the extent to which the firm is permitted to pass onto consumers any specific cost increases in the form of higher prices. He finds that when the firm has increasing costs and is allowed to pass onto consumers a proportion of the costs increase that is sufficient to maintain expected profits, then the associated level of reliability will be increased. This, effectively, would resemble a situation of rate-of-return regulation. However, if the firm is forced to absorb the cost increase to the detriment of its level of expected profits, then the firm's response will be to minimise the loss of expected profits by lowering reliability. Thus, if consumers are protected against the cost increase, this protection will be at the expense of lower reliability. Fraser's conclusions are important to consider in a regulatory setting. Often, regulators impose a price-cap with a gradual price decrease (through the X-factor) reflecting the regulator's expected improvement in efficiency. Fraser's results imply that if the X-factor is set too high and the firm cannot achieve the regulatory cost targets, its strategic response to maintain sufficient profits will be to lower reliability. To solve this problem, Fraser proposes to include a reliability element into the price-cap. This benefits consumers in a situation where the firm is required to absorb a cost increase because it can no longer protect profits by reducing its existing reliability level.

The effects of a regulatory shift from rate-of-return regulation to price-cap regulation are studied by Kidokoro (2002). He finds that if quality is capital driven, the regulatory shift reduces both price and quality. In the reverse case, when quality is effort-related, the shift to price-cap regulation will reduce price and raise quality. Kidokoro argues that price-cap regulation is not suitable to regulated industries in which the amount of investment is crucial for the resulting level of quality. In this case, some rate-of-return regulation amendments may increase the total social surplus. In other words, a hybrid form of price-cap regulation and rate-of-return regulation may enhance social surplus. When applied to the electricity network industry, Kidokoro's results suggest that a shift to price-cap regulation will cause quality degradation as quality (reliability) is predominantly capital-related. For effort-related quality, a reverse tendency may be expected. This suggests that it is preferable to price-cap regulate only those activities that are primarily effort-related while maintain a rate-of-return regulation system for capital-related activities.

In line with the previous authors, Weisman (2002) finds that under price-cap regulation, the regulated firm's incentive to invest in service quality increases with the level of the price-cap. Secondly, Weisman shows that the incentive to reduce investment in quality under price-cap regulation may be tempered by the regulated firm's participation in complementary, competitive markets. A reputation of poor quality in the provision of monopoly services can spill over to adversely affect sales in the competitive markets the firm is also engaged in.[22] Weisman also analyses the effect of increased information dissemination actions, i.e. exposing the firm's performance to the public. He finds that exposing performance provides the regulated firm with incentives to increase investment in quality without distorting the efficient investment in cost-reducing effort.

Anecdotal evidence points out the problems of quality under stricter price regulation. Weisman (2002) provides the examples of Idaho and Oregon where an incentive system for the telecommunications industry was abandoned due to problems with service quality. According to Ter-Martirosyan (2003), the State Public Utility Commission of Oregon terminated performance-based regulation plans for Pacific Power in 1995 because of the resulting low quality of service, and reintroduced it in 1998 only after incorporating strict quality standards for reliability. In Hungary, a large increase in the number of interruptions was noticed after the introduction of price-cap regulation of electricity distribution firms and this accelerated the introduction of explicit quality regulation schemes (Tersztyanszky 2003). These examples tend to confirm the concern with quality decline under price-cap regulation. Unfortunately, there are only a few studies available that collected empirical evidence of quality effects under price-cap regulation. For the telecom industry, Sappington (2003) compares four empirical studies on the effects of incentive regulation on quality to retail telephony consumers in the US.[23] He finds that these studies produce diverging results; the hypothesis that quality declines with stricter regulation cannot be unambiguously proved for the case of US telecom. This, as he argues, reflects in part the limited success of the existing studies in capturing all relevant aspects of regulatory policy.

For the electricity industry, as far as known, there is only one empirical study by Ter-Martirosyan (2003) that looks at the effects of price-cap regulation on reliability.[24] She found that price-cap regulation indeed led to worse quality performance in terms of an increase in the SAIFI and CAIDI indicators. In her study, she analyses a pooled sample of 78 electricity firms from 23 states of the US during the period 1993 to 1999.[25] The econometric model takes into account the type of regulatory regime, the presence of quality standards, the per capita income in the state, average length of line per consumer, the share of underground lines, the share of self-generation by the firm, and the damage caused by severe weather conditions in the territory served by the firm. Her analysis supports the hypothesis that price-cap regulation has a negative impact on quality if no precautions are taken to safeguard quality. This is particularly true for the duration of interruptions. However, incorporation of quality standards appears to reduce this effect. Then, interruption duration remains the same or even improves in some instances. In contrast, price-caps do not seem to have significant impact on the frequency of interruptions. Possible explanations given for this are the absence of data and the different causes that contribute to frequency of interruptions. She refers to a study conducted by the Oregon Public Utility Commission showing that the main cause of interruption occurrence is equipment failure, whereas interruption duration is mostly affected by storms and the time it takes to repair the damage. Thus, she argues, if price-cap regulation affects the cost structure of the firm, the impact on equipment is a long-term effect and related changes in reliability may not be noticed in the short run. The appropriate model to be used for interruption frequency would be one with lagged values of regulatory regimes. The short history of incentive regulation, however, does not allow the testing of such models yet.

Ter-Martirosyan also considers the potential problem that price-cap regulation and quality standards may be endogenous: They are more likely to be applied when the firm has a poor performance. After adjusting her model to capture this effect, she finds that price-cap regulation still negatively affects reliability performance and that quality standards help to mitigate this problem. Another issue studied by Ter-Martirosyan is how incentive regulation affects the firm's spending on operational and maintenance. In principle, incentive regulation is designed to promote efficiency and therewith reduce spending levels. But, as discussed at length in previous sections, it also comes with the risk of adverse cost reductions at the expense of quality. Ter-Martirosyan analyses reductions in spending for firms with and without quality standards. In line with expectations, she finds that in both cases, spending levels have fallen which suggests that the firms have operated more efficiently. However, she finds that firms with quality standards have reduced costs less than those without. In the former case, spending levels were reduced by 17 percent since 1993 while for firms without quality standards this decrease has been about 37 percent. This is a substantial difference, which seems to support the hypothesis that price-cap regulation without quality measures generates perverse incentives to under-spend on quality. In conclusion, Ter-Martirosyan's empirical study proofs that the problem of quality degradation under price-cap regulation is a real one and needs to be dealt with appropriately.

2.6 Conclusions

2.6.1 Synthesis

This chapter has investigated the underlying arguments for regulating (natural) monopolies. As widely argued in the literature and supported by empirical studies, monopoly leads to allocative and productive inefficiencies as well as suboptimal quality levels. At the same time, the justifiable financial interests of the firm need to be taken into account. Generally speaking, the objectives of regulation are to promote high economic efficiency and optimal quality but at the same time ensure the firm's viability.

There are generally speaking two systems of price regulation, rate-of-return regulation and price-cap regulation; each has a different impact on quality in electricity distribution. Under rate-of-return regulation, the firm has a tendency to overcapitalise as the rate-of-return is typically set higher than its costs of capital. This leads to overcapitalisation and consequently to high quality levels. As the price regulation system automatically leads to high quality, there is no explicit need for the regulator to deal with quality. In the words of Kahn (1970), "the responsibility for quality issues remains with the supplying companies". Under rate-of-return regulation, regulators can thus indirectly escape the quality regulation problem, but this comes at the cost of lower efficiency and suboptimal quality. Price-cap regulation provides strong incentives to operate at higher productivity levels. This incentive can, however, adversely result in cost reductions at the expense of quality. Both theory and empirical evidence indicate that this is an important problem and requires the regulator to install additional quality measures to counteract this problem. If regulators do not manage to deal with quality issues properly, consumers may effectively be worse off than under rate-of-return regulation.

The strong modern focus on incentives and efficiency in price regulation now create an explicit need for quality regulation. At the same time, however, effective quality regulation is complex to design, implement and administer. These complexities become manifest under stricter price regulation schemes. This suggests that at some point the advantages of stricter price regulation – resulting in enhanced efficiency levels – will not outweigh the additional regulatory costs of installing and maintaining adequate quality regulation. This implied trade-off between price and quality regulation takes place in two ways. Firstly, it depends on how the regulator designs the price-cap. Generally, a mixture of price-cap systems can be identified with some regulators applying strong efficiency incentives and other regulators adopting a more lenient system that more or less resembles rate-of-return regulation. An adverse impact on quality is more likely to be observed in the former case than in the latter one. Secondly, the trade-off between price and quality regulation depends on the extent by which the regulator is able to integrate quality into the price-cap. In doing so, he faces a number of problems and complications. An effective integrated price-quality regulatory system requires a proper balance between efficiency properties (the price aspect) and a proper mechanism to incorporate quality (the quality aspect). Not only higher productivity matters, but an appropriate quality level is an important regulatory objective as well. Higher effectiveness at the price side requires a proportional increase in effectiveness at the quality side. The regulator's ability to successfully deal with these complexities will certainly contribute to a more effective integrated price-quality regulatory system.

2.6.2 The Way Forward

The next two chapters address the price and quality aspects of regulation, respectively. Chapter three analyses the different price-cap approaches and their interaction with quality. Chapter four deals with the problems experienced in establishing quality regulation and integrating quality into the price-cap.

Notes

[1] Cit. by Newbery (1999, p. 28).

[2] Next to market power, other sources of market failure are asymmetric information and externalities (Armstrong et al. 1994, p. 11).

[3] Decreasing average costs means that $AC_x<0$ and thus $\partial / \partial x \cdot (FC+VC)/x=VC_x/x-[FC+VC]/x^2 <0$ which implies $VC_x - (FC+VC)/x<0$ thus $VC_x<AC$. Or in words: if an additional unit of production decreases average costs, then the costs of that unit are necessarily lower than the costs of the previous units i.e. marginal costs are decreasing.

[4] As will be shown later, an important critique on traditional regulation has been its weak productive efficiency properties as was exposed firstly by Averch and Johnson (1962).

[5] Nobel Laureate George Akerlof's seminal "Market for Lemons" paper (Akerlof, 1970) is one of the pioneering studies of the role of quality in markets with imperfect information. As he shows, information asymmetry between firms and buyers can result in market failure. See also Dorfman and Steiner (1954), Dixit (1979), Levhari and Peles (1973) and Leland (1979).

[6] The Spence and Sheshinski models are very similar and were published around the same period. They were developed independently from each other (Sheshinski 2003).

[7] The reader is referred to Spence (1975) and Sheshinski (1976) for the original analysis. Extended analyses based on the Spence-Sheshinski models include Besanko and Donnenfield (1988) who consider the case of product variety. Lambertini (1998) studies a multiproduct monopolist supplying vertically differentiated varieties of the same good. Sibley (2002) makes demand dependent on consumer disposition to the firm in addition to quantity and quality of the good.

[8] See for example Brealy and Myers (2000).

[9] Beta's are a common statistic used by investors and measure the extent to which the firm's returns vary relative to those of a diversified portfolio of equity holdings. The beta indicates whether an investor with a diversified portfolio would take on more risk by investing in a particular firm. The higher the beta, the larger the increase in riskiness of the investor's portfolio. See also Brealy and Meyers (2000) pp. 195-203.

[10] The regulator's inferior information position also may give rise to a host of strategic manoeuvres by the firm. For example, firms usually claim that the regulator's estimation of the X-factor is too high, not realistic, and if sustained, will definitely lead to bankruptcy. This, for example, happened when X-factors were set for Dutch electricity distribution firms in 2000 (TZH 2000).

[11] Demsetz (1968) already acknowledged that given the problem of uncertainty in combination with a long-term contract, if the firm who won the bid enjoys windfall profits for a longer period, it may be

desirable to employ a costs-plus regulatory scheme or enter a clause that reserves the right to renegotiate the contract. Such an approach would already resemble regulation.

[12] Rate-of-return regulation is synonymously labelled "costs of service" or "costs-plus" regulation.

[13] See for example Train (1991) for a detailed discussion of the Averch-Johnson results.

[14] In order to sustain the incentive, the downward adjustment of prices should not effectively expropriate the cumulated profits resulting from the improvements during the last regulatory period.

[15] Subsequent articles have echoed and confirmed the advantage of fixing the regulatory lag. Train (1991) mentions Bailey and Coleman (1971), Davis (1973), Klevorick (1973), Bawa and Sibley (1980) and Logan et al. (1989).

[16] As also argued in paragraph 2.4.1, the regulator will set the fair rate-of-return at a level equal or higher than its capital cost in order for the firm to at least break even. In practice however, the fair rate-of-return may be lower than the firm's capital cost. For example, in the former communist countries, political factors often led prices to be below actual costs. In this case, an undersupply of quality is more likely to occur. This hypothesis is supported by the general observation of historically lower investment and quality levels in electricity infrastructures in these transitional countries compared to Western Europe or the US.

[17] The converse is also true: If quality is labour-using, then rate-of-return regulation is likely to exacerbate the quality problem.

[18] The issue of optimal quality is explored in detail in chapter four.

[19] An overview and discussion of different regulatory incentive schemes can for example be found in Train (1991), Armstrong et al. (1994) and Liston (1993). A comprehensive theoretical assessment of incentive regulation is contained in Laffont and Tirole (1993).

[20] Baron and Myerson (1982) model this by defining social surplus as a weighted sum of consumer and producer surplus with a higher weight attached to the former.

[21] For an overview, see Laffont and Tirole (1993).

[22] Such spill-over effects have played an important role in the improvement of reliability levels in Italy. See also Annex III for the Italian case study.

[23] The studies that are compared are those by Ai and Sappington (1998), Roycroft and Garcia-Murillo (2000), Banerjee (2003) and Clements (2004).

[24] Ter-Martirosyan (2003) uses the SAIFI and CAIDI indicators to measure reliability. See also Annex I for an overview of reliability indicators.

[25] One may speculate that the reason why quality degradation occurs in electricity and not in telecom is that given the rate of technological advance in the latter, quality being constant would effectively be a decrease in quality. Vickers and Yarrow (1988) note for example that "*British Telecom's quality of service has not deteriorated since privatisation, but it has not improved much either. Given the rate of advance of telecommunications technology, this record is poor.*"

3

Price-Cap Regulation

3.1 Introduction

3.1.1 Background

Presently, price-cap regulation is widely being applied in network industries such as electricity, gas, telecom and water. The main advantage of a price-cap system lies in the strong incentives it generates for higher productive efficiency. Price-caps unlink prices from actual costs by imposing a predefined change in prices over the course of a fixed regulatory period. The annual change in prices is determined by the X-factor. If the firm manages to reduce its costs in excess of the X-factor, it earns additional profits and conversely, if it performs worse than the X-factor, it earns less profit. This is the basic incentive provided by the price-cap system.

As was established in the previous chapter, price-cap regulation in principle leads to quality problems as there is a risk that the implied cost savings are achieved through undesired quality reductions. The severity of this quality problem will depend on the choice of the price-cap approach. Consequently, the preferred way to deal with the quality problem will also vary as a function of the price-cap approach. The objective of this chapter is to develop a taxonomy of price-cap approaches and explore alternatives for integrating quality into the price-cap.

3.1.2 Chapter Outline

This chapter is structured as follow. Section two starts with a general introduction to the topic of price-cap regulation. Here, the elements of the price-cap and the role of benchmarking are explored. Section three develops four strategies for implementing a price-cap. Section four identifies two main price-cap approaches and evaluates these with respect to their quality incentive properties. Then, proposals to integrate quality into the price-cap are developed.

3.2 Price-Cap Elements

3.2.1 Price-Cap Formula

Economic theory predicts that maximum efficiency – both in the allocative and productive sense – is achieved under perfect competition.[1] One of the main features of a competitive market is that no single firm can influence the going market price. Each firm's profit is then, amongst others, determined by the extent to which this firm is able to operate more efficiently than its competitors. In the context of regulating monopolies, similar incentives can be created by setting the allowed price on an exogenous basis i.e. independently from actually incurred costs. Given that prices are fixed, ceteris paribus, operating at higher productivity levels i.e. producing the same level of outputs at lower costs will drive up the firm's profits.[2]

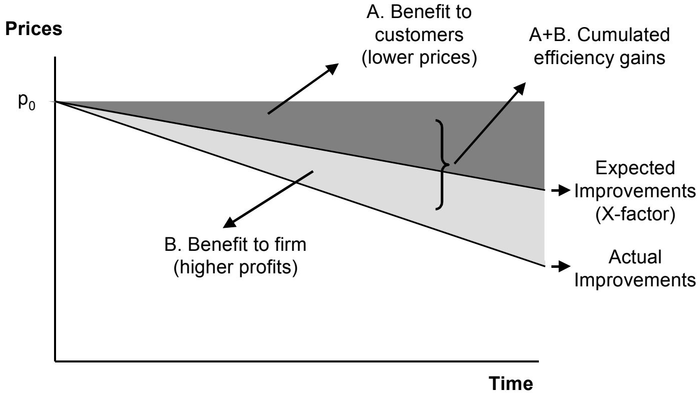


Figure 3-1. Simplified representation of the incentives provided by the price-cap system. Consumers enjoy gains (represented by (area A) due to a reduction in the initial price P_0. The firm retains extra profits due to cost savings in excess of the X-factor (area B). For society as a whole, efficiency savings are given by the area A+B.

The main difference between price-cap regulation and traditional rate-of-return regulation is that under the former system, prices are no longer directly based on the firm's actual costs. At the one extreme, under a pure rate-of-return scheme, prices would be set on the basis of the firm's actual costs. This provides no incentives for higher productivity. The other extreme is to completely unlink prices from actual costs; this provides very strong incentives for productivity improvement. Price-cap systems are located somewhere between these two extremes. That is, prices and costs are detached from each other, but not to the full extent; there still remains some interdependency. In practice, the regulator sets prices not on the basis of the firm's actually incurred costs, but rather on a level of cost that the regulator considers efficient. The difference between actual costs and the regulatory estimation of efficient costs is reflected in the X-factor. This X-factor applies for a given number of years (the regulatory period) and determines the annual change in prices in such a way that prices move in line with the anticipated efficiency improvements. Through the X-factor, consumers directly participate in the expected cost reductions in the form of a lower price.[3] On the other hand, the firm also benefits in the case that it manages to reduce its costs in excess of the X-factor. The residual cost savings can then be retained in the form of higher profits.

The length of the regulatory period and the level of the X-factor are the two main variables in the price-cap system. Typically, prices are also adjusted for inflation in recognition of the fact that the cost of goods and services used in the production process will change over time and that this change in price level is generally not controllable by the firm. In its most general form, the price-cap formula is then given by:

$$p_t = p_0 \cdot (1 + CPI_t - X)^t \tag{3-1}$$

Here, p_0 is the initial price, p_t is price for year t of the regulatory period, CPI is the consumer price index, and X is the annual price adjustment. By limiting the duration of the regulatory period, the regulator can make sure that differences between actual productivity improvements and anticipated improvements are retained only for a fixed period. In practice, a regulatory period of between three and five years is deemed a reasonable compromise (KEMA 2004). The inflation factor is typically the one published by statistical institutions and can, apart from the CPI for example be the retail price index (RPI), or producer price index (PPI), or a combination of these or other inflation indices.[4]

One way to look at the price-cap is as a contract between the regulator and the firm. The regulator sets the X-factor at the start of the regulatory period and does not change this afterwards. With price levels given, a stable environment is created where the firm can engage into activities to improve productivity. All improvements in excess of the X-factor translate into profits, which the firm can keep. If the regulator underestimates the improvement potential of the firm, these excess profits can grow quite large and in turn lead to political or social pressure to claw back some of these profits. This, however, would imply a breach in the regulatory contract, which may create a credibility problem and potentially destabilise the price-cap system. Firms may no longer consider the regulator committed to its initial promise to let them retain excess profits in case they achieve extraordinary performance. Anticipating future claw-backs, they will be less motivated to achieve any improvements as this would no longer provide any additional benefits. Regulatory commitment thus is an important condition for assuring an effective price-cap system.

The risk of ex post profit claw-backs, roughly speaking, decreases with the regulatory ability to set an appropriate X-factor. If the regulator underestimates the firm's true improvement potential, the X-factor would be too low and the firm would earn excessive profits – giving rise to distributional problems. Similarly, overestimation of the X-factor would jeopardise the financial sustainability of the firm, as the X-factor would impose unrealistic improvement targets. A related problem is that the firm may be able to materialise its improvement potential only over some time. For example, even if the firm is initially overstaffed, contractual constraints may prevent a quick dismissal of personnel. This may take a number of years to accomplish (e.g. through natural outflows) or would come at high costs.

If the regulator is able to accurately predict the firm's future productivity improvements, it could set the X-factor on this basis. Then, the firm would not earn too high excess profits while at the same time, financial sustainability of the firm is also assured. A better assessment of the firm's true productivity improvement potential can thus lead to a better balance between interests of the firm and consumers. In summary, the X-factor should be low enough to leave the firm with sufficient funds and it should be high enough so that consumers can also share in ongoing productivity gains. It will be evident, however, that quantifying the productivity potential, and therefore setting the X-factor, is seriously complicated by the regulator's poor informational position relative to the firm.

Generally speaking, one may assume the firm to have private (albeit incomplete) information about whether and how much it could improve its efficiency. This information is not available to the regulator and consequently, the regulator cannot determine the most appropriate X-factor. Furthermore, the firm could strategically exploit its superior informational position by talking down the X-factor – claiming that it is based on inaccurate estimations and provides unrealistic or unattainable targets. Clearly, the regulator's ability to assess the firm's true productivity improvement potential can greatly benefit the effectiveness of the price-cap system. Benchmarking analysis can play an important role in this regard.

3.2.2 Benchmarking

Higher productive efficiency, or higher productivity, implies that firms produce the same level of outputs but using fewer inputs (or more outputs using the same level of inputs). Ideally, a firm should operate at the highest possible productivity level. Then, the firm is said to be operating at the productivity frontier (Coelli et. al 1998, p. 3). If the firm is not located on the frontier, it is operating inefficiently i.e. there is scope for efficiency improvement. This is demonstrated in

Figure 3-2. The curve OF' represents the production frontier i.e. the maximum output attainable from each input level. Hence, it reflects the current state of technology in the industry. Firms can operate either on that frontier if they are efficient or beneath the frontier if they are inefficient. Point A represents an inefficient point whereas points B and C represent efficient points. If a firm were to operate at point A, it would be classified as inefficient because it could well increase output to the level associated with point B without requiring more input. Alternatively, it could produce the same level of output using less input i.e. produce at point C on the frontier.

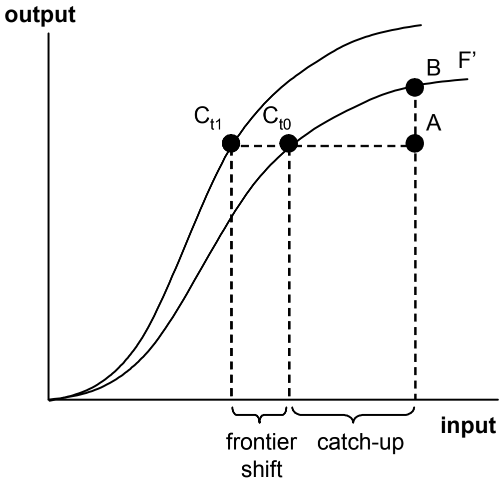


Figure 3-2. Simple input and output example. Firms B and C are efficient, these are located on the productivity frontier. Firm A is not located on the productivity frontier and is inefficient. Over time, the frontier will shift as firm C improves productivity further.

The X-factor should ideally incorporate the productivity improvement potential of the firm. This could be done by measuring the distance between the firm's productivity level and that of its projection on the frontier. However, the regulator should also take into account that the productivity frontier itself may change over time. Firms that are initially fully efficient i.e. are located on the frontier, could increase their productivity even further over time as a result of technological change. This leads to a shift of the productivity frontier; this effect is known as the "frontier shift". In setting the X-factor, not only initial static inefficiencies, but also the dynamic improvements in productivity would need to be considered. Thus, benchmarking would entail two separate efforts. Firstly, the measurement of the so-called catch-up effect which is the closing in on the initial gap between the current (inefficient) productivity level and the frontier. Secondly, the frontier shift which reflects the ongoing improvement in productivity and which even firms located on the frontier would be able to realise ("running to stand still"). Ideally, both the catch-up effect and frontier shift should be measured and included in the determination of the X-factor.

In practice, the frontier shift should be forecasted as it can only be measured once the technological changes have actually been achieved. Experience, however, shows that the X-factor usually does not explicitly consider the frontier shift. The probable reason for this is that the frontier shift effect tends to be rather small in comparison with the catch-up effect.[5] In contrast, the initial closing in on the frontier (the catch-up effect) has received much more regulatory attention. For measuring this initial efficiency improvement potential, regulators often make use of benchmarking analysis (Jamasb and Pollit 2000). Benchmarking, as the name suggests, is based on the concept of comparing the performance of the firm to that of best practice. Firms that operate at the productivity frontier act as the benchmark for firms that are not yet located on this frontier. The frontier firms (or peers) operate at maximum productivity levels and, by definition, have an efficiency score of 100 percent. For other firms, the efficiency score is measured as the distance to the frontier. The further away from the frontier, the lower the efficiency score will be.

Table 3-1. Overview of benchmarking techniques used by energy regulators (either explicitly or implicitly). Sources: Jamasb and Pollit (2000), Sumicsid (2003b), and personal communications with regulators. See Annex II for a description of the different benchmarking techniques.

	Data Envelopment Analysis (DEA)	Total Factor Productivity (TFP)	Parametric Techniques	Partial Indicators	Value Chain Method	Norm Models
Austria	X					X
Australia (NSW)	X	X	X			
Australia (Queensland)	X					
Australia (Victoria)				X		
Brazil	X					
Canada (Ontario)		X				
Chile						X
Columbia	X					
Denmark	X					
Finland	X					
Hungary	X					
Malaysia	X		X	X		
Netherlands	X	X				
Northern Ireland	X					
Norway	X				X	
Singapore	X		X			
Slovenia	X		X			
Spain						X
Sweden	X		X			X
UK	X	X	X			

An important advantage of benchmarking is that it provides information on the basis of empirical data; in principle, all firms should be able to operate equally efficient as their peers.[6] Benchmarking analysis can thus provide the regulator with valuable information that can be used for setting the X-factor. However, the validity of the benchmarking analysis will be driven by the way how the frontier and consequently efficiency scores are measured. In the context of electricity distribution, firms use different inputs (capital, labour) to provide different outputs (or services) to consumers (connections, energy, quality, etc.). While all firms use broadly the same type of inputs, some providers may use proportionately more of some inputs and less of others. The mix of inputs used depends upon, among other things, management practices and the operating environment. Similarly, the nature of services provided by networks varies according to the nature of consumer demands. For example, some firms may need to maintain significant network capacity to distribute electricity to a small number of consumers while others may serve a large number of consumers with a highly variable demand. Furthermore, there may be other factors such as climate, geography, or demography that influence the firm's costs. In the calculation of productivity, the multi-dimensional nature of the production process as well as the presence of structural differences between firms should be taken into account. There are a number of benchmarking techniques that can be used for this purpose. An overview and discussion of the different benchmarking techniques is provided in Annex II.

3.3 Price-Cap Strategies

3.3.1 Classification

As may be observed in Table 3-1, benchmarking is an important regulatory instrument to identify the scope for productivity improvement and to consequently set the X-factor. However, there are different ways to translate the results of the benchmarking analysis (the efficiency score) into the X-factor. One extreme would be to directly link the X-factor to the efficiency score. In this case, the regulator could perform a benchmarking analysis at the start of the regulatory period and set the X-factor for each firm based on its efficiency score. This efficiency score represents the extent by which the firm could reduce its costs down to the level of what would be considered efficient. The X-factor then imposes a gradual price reduction from the initial price towards a price that reflects an efficient level of cost (including a reasonable profit). If n is the duration of the regulatory period in years and θ is the efficiency score obtained from the benchmarking analysis, the X-factor for a given firm would be set such that:

$$(1-X)^n = \theta \tag{3-2}$$

The firm thus needs to reduce its costs in line or in excess with the X-factor in order to maintain a high level of profitability. Furthermore, in the case of multiple-firm regulation, the efficiency score θ and therefore the X-factor would reflect efficiency improvement potential relative to other firms. This introduces a degree of competitive pressure: Firms that operate at higher productivity levels would obtain a higher efficiency score and consequently get a lower X-factor.

However, the link between the benchmarking analysis and the X-factor can also be indirect. If the regulator feels that it can only imperfectly perform a benchmarking analysis, he may wish to use the efficiency score as a starting point for setting the X-factor rather than imposing a mechanistic link between the X-factor and the efficiency score. The benchmarking results would provide information on the range where the X-factor could be located. This information can then be used as input for the decision on the X-factor.

For now, an implicit assumption was that the regulator set the X-factor at the start of the regulatory period. Alternatively, the regulator could choose to set the X-factor at the end of the period i.e. once the firm has realised its productivity improvements. This in fact takes away any uncertainty in the level of the X-factor, as it would then be based on actually achieved improvements. However, this approach has the disadvantage that the firm has no incentive to achieve productivity improvements in the first place as it may anticipate that these will be clawed back – similar as under a system of rate-of-return regulation. To avoid this problem, the regulator can impose limits to the level of the X-factor. For example, the regulator could set an initial X-factor and only adjust this if the firm's profits move outside some predetermined range. The firm thus always retains part (or potentially all) of its realised improvements. In the case that there are multiple firms being regulated, the regulator could set the X-factor on the basis of actually observed changes in the average performance of all firms. This introduces competitive pressure, as firms that improve beyond the average would enjoy higher profits than those who perform less than average.

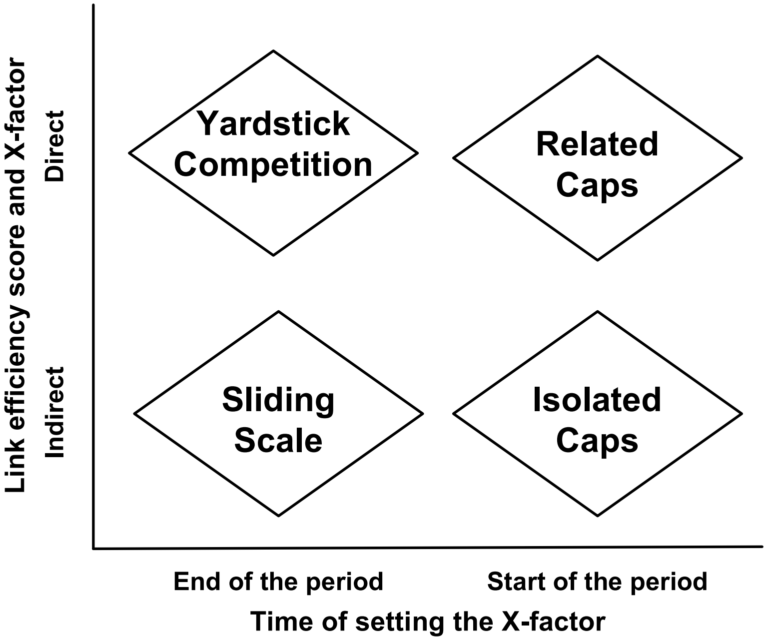


Figure 3-3. Classification of price-cap strategies.

In summary, two dimensions can be identified in which the process of setting the X-factor can be characterised. Firstly, as shown on the vertical axis in Figure 3-3, the regulator can choose for either a strong direct coupling or a loose indirect coupling between the benchmarking analysis and the X-factor. The second dimension, as shown on the horizontal axis in Figure 3-3, is concerned with the timing of the X-factor setting in relation to the benchmarking analysis. The X-factor can be set at the same time as the benchmarking analysis i.e. at the start of the regulatory period, or afterwards i.e. at the end of the regulatory period. Figure 3-3 shows how combining the options in each of the two dimensions leads to four possible strategies for setting the X-factor. In the following sections, these strategies are explored in more detail.

3.3.2 Yardstick Competition

Yardstick competition introduces a strong competitive aspect to the process of setting the X-factor. In Shleifer´s (1985) original definition of yardstick competition, the price for each firm is set equal to the average cost of all other firms in the regulated industry. There are some variations on this theme. For example, the price can be set on the basis of the average cost of all firms (including the firm under consideration), or one could apply some weighted average of costs to calculate the yardstick price. Irrespective of the specific formulation, the main idea is that the firm's profitability is no longer determined only by its own cost performance but driven by how well it manages to reduce costs relative to others. This gives a strong incentive to increase performance – similar to the incentive observed in competitive markets. If a firm manages to reduce its costs more than the yardstick, it will earn a higher profit and conversely, firms that lag behind average performance will earn less profits and possibly even incur losses. As all firms have an incentive to reduce costs, this also brings down the average costs of the industry. Thus, a continuous downward adjustment of the prices would take place where each firm's effort to reduce costs in excess of the average simultaneously leads to a decrease in the yardstick itself.

In the price-cap context, the X-factor under a yardstick competition scheme would be set on the basis of actual improvements in productivity. Thus, there is in principle no need for the regulator to make any predictions about productivity improvement potential as this information would be automatically revealed through the yardstick scheme. Also, as prices continuously track realised improvements over time, efficiency gains are quickly transferred to consumers. In essence, under yardstick competition the regulator would no longer have to set the X-factor but would simply adjust prices each time on the basis of some index of average cost.[7]

In his seminal paper on yardstick competition, Shleifer (1985) noted that an important aspect of measuring the yardstick is the need to adjust for possible structural differences between firms. Setting prices on the basis of average costs suggests that firms are perfectly comparable to one another. This may not necessarily be true as there may be structural differences in the operating environment across firms. Some firms may have to deal with specific factors, which lead them to incur relatively higher costs than others. Furthermore, one also needs to take into account the multi-dimensional nature of the firm's production process. There may be more than a single input or output factor involved in providing the distribution service. Neglecting such factors in the determination of the yardstick would disadvantage some firms and provide other with an unintended advantage. To deal with this problem, more sophisticated notions of average costs could be used. The use of benchmarking methods, which incorporate multiple input and output factors and allow corrections for structural differences, can play an important role in this process (DTE 1999).

In addition to the comparability problem, there are two other main problems attached to yardstick competition, namely commitment and collusion (Weyman-Jones 1995). The collusion problem is related to the fact that the firms may strategically cooperate to influence the outcome of the yardstick system. For example, firms may collectively report higher costs than actually incurred in order to drive up the yardstick. The scope for collusion increases as the number of firms is smaller. Therefore, in order for yardstick competition to be effective, a large number of participating firms is a necessary (but not sufficient) condition.

The third problem is that of regulatory commitment. Yardstick competition assumes that the regulator is committed to the regulatory contract. This means that, irrespective of the outcome, the process by which the yardstick is calculated is not changed afterwards. In principle, this should also hold in the case of bankruptcy of one or more of the participating firms. Similar as in a competitive environment, firms who perform better earn exceptional profits while others that lag behind, earn less profit, potentially become unprofitable and eventually may go bankrupt. If the yardstick system is to remain credible, bankruptcy of one or more firms should not be excluded as a potential outcome, implying that the regulator should not adjust the rules of the system ex post to prevent ill-performing firms from going bankrupt. As pointed out already in chapter two, bankruptcy of an electricity distribution firm has substantial social and therewith political impact. It therefore remains questionable if such firms would in practice be allowed to go bankrupt.[8]

3.3.3 Related Caps

For yardstick competition to be fair, all firms should have the same initial scope for improvement. If this is not the case, then firms who are initially less productive than others could reduce costs more than others and subsequently drive down the yardstick. These firms would then consequently earn higher profits than firms with less initial scope for improvement. However, these profits would be the result of an unequal starting position and therefore not be conceived as fair by the other firms. To deal with this problem, one could assure that firms are first brought to the same productivity level. Creating this level playing field is the basic idea of the related caps strategy. The related caps strategy may thus be considered as the preparatory phase before moving to yardstick competition.

Under related caps, the regulator would set the X-factor at the start of the regulatory period on the basis of an assessment of the relative efficiency of each firm. Clearly, similar as under yardstick competition, the related caps strategy can only be applied in case of multiple firm regulation. Each firm would be allowed a different price and X-factor, reflecting its starting productivity level and it improvement potential, respectively. The X-factor would be directly driven by the results of the benchmarking analysis. Hence, there will be a strong degree of interconnection between the X-factors and prices for different firms. The ability to compare firms in a proper way i.e. account for the multi-dimensional nature of the distribution service and incorporate structural (accounting) differences between firms is therefore an important precondition for the related caps strategy to be effective.

Box 3-1. The problems of related caps in the Netherlands. Taken from Ajodhia et al. (2003)

DTE published its first decision on the X-factors for electricity distribution networks in September 2000. These X-factors were strongly driven by the results of a DEA benchmark report. The DEA benchmark was applied to the sample of 20 Dutch distribution firms. As an input factor for the benchmarking, DTE chose the total cost, which is the sum of operating expenditure, depreciation, and a standardised return on assets. In order to harmonise depreciation and book value data across firms, DTE performed a backward calculation of book and depreciation values. In doing so, however, a number of assumptions and approximations had to be made. Due to the lack of detailed data, the standardisation was performed on an aggregate basis, thereby ignoring the differences in lifetime and age across asset categories. Also, as historical investment profiles were not available, a virtual annual investment profile was assumed when recalculating the asset and depreciation values.

The September 2000 decisions on the X-factors led to a wave of protest and formal appeals by the industry. The main critique was aimed at the use of benchmarking as a way to set tariffs: Efficiency scores from the DEA analysis were mechanically translated into X-factors. The result of this was that flawed data – in particular due to the standardisation of capital costs – could lead to wrong efficiency scores and in turn, to wrong X-factors. As the efficiency score of each company was in principle linked to that of the others, so were the X-factors and therefore also the prices. Obviously, companies were not comfortable with the idea that their X-factor and thus allowed income would be driven by data errors. Additionally, the fact that DTE widely published the benchmarking results did not help in this regard. As a result, the relationship between regulator and industry became increasingly hostile. On the one hand, DTE confirmed its decisions; on the other hand, the network companies refused to accept the – in their eyes unjust and erroneous – X-factor decisions.

At some point however, DTE had to revise its initial decisions in September 2001; the main difference with the initial decisions was an increase in the quality of data. An independent audit was performed to verify and improve the output factor data, while the CAPEX standardisation was refined by considering each individual asset and the actual historical investment profile. The data improvements led to higher efficiency scores and to lower X-factors. However, the companies' main critique points were still not thoroughly met, and there still remained problems with the data. DTE responded to this by initiating a special project with the objective to remove any remaining data problems. As a result, a second revision of the benchmark analysis and X-factors was published in August 2002, but this did not prevent the network companies from confirming their appeals, as they did not consider DTE's corrections to be sufficient. Eventually, in October 2002, the Courts overruled the X-factor decisions. However, the motivation for this decision was the fact that according to the Dutch Electricity Act, DTE should have applied a uniform X-factor (instead of an individual X-factor for each company) in the first place.

3.3.4 Isolated Caps

Under yardstick competition and related caps, there is a direct link between the efficiency score and the X-factor. As efficiency scores directly feed into the X-factors, any errors in the efficiency scores will also affect the X-factors, the price, and eventually the profitability of the firms. Errors in the efficiency score can be caused by model errors and/or data errors. Firstly, model errors are concerned with invalid model specifications e.g. exclusion of relevant variables (input and output factors) or do not allow correction for structural differences. Secondly, data errors refer to the use of erroneous data, for which a variety of causes may play a role. Clearly, one would like to avoid data errors as well as model errors to drive the X-factors. If initially a benchmarking analysis is conducted and results found to be wrong, the analysis may have to be rerun and the X-factor reset. If such adjustments often take place, the credibility of the system suffers. This is particularly problematic given that the X-factors are interrelated: Errors in the efficiency score of one firm can potentially influence the X-factors of other firms. This feature also makes the system vulnerable for strategic data reporting and submissions – in particular when the number of firms is small.

In the case that there is only a single firm to be regulated, or if the regulator considers the yardstick competition or related caps strategies not feasible, the isolated caps strategy may be considered. Here, the regulator sets the X-factor for each firm on an individual basis at the start of the regulatory period. For this purpose, the regulator ma y still make use of benchmarking analysis but the link between the efficiency score and the X-factor would not be direct. The benchmarking results are used as an indication of inefficiency and would only indirectly influence the X-factor. This has the advantage of reducing the sensitivity for data or modelling errors. Each firm would here be considered in isolation even though the benchmarking analysis may be applied to all firms together.[9]

3.3.5 Sliding Scales

It may be that, for some reason, the regulator cannot perform benchmarking analysis or considers its results of limited use in setting the X-factor. Lack of information about the firm's true productivity improvement potential may, as discussed earlier, lead to two basic problems. On the one side, the X-factor may be set too low and the firm will earn excessive profits. On the other side, the X-factor may be set too high; this can cause financial problems for the firm. Taking this into account, the regulator could decide to adjust the X-factor in such a way that the firm's profit varies only within a given range. Under this strategy, which is known as sliding scale, the regulator sets the X-factor as a function of the profitability of the firm (e.g. as measured in terms of its rate-of-return). If, at the end of the regulatory period, the firm's profit exceeds some predetermined band, the X-factor is adjusted such that profits are brought back within this band. Conversely, if actual profits are higher than the allowed maximum, the X-factor is adjusted in such a way that these profits are reduced down to the level of the maximum. A similar procedure would also apply for the minimum profit level. In between the two extremes, the X-factor would not be adjusted i.e. the firm would earn the rate-of-return as observed at the end of the regulatory period.

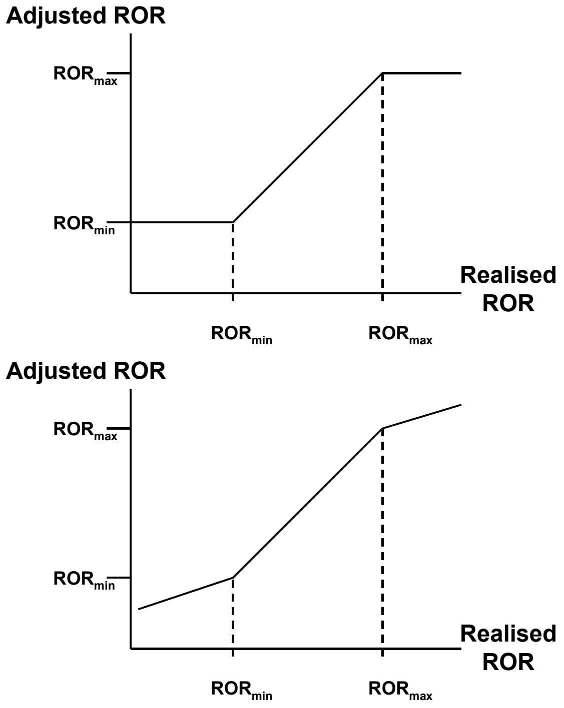


Figure 3-4. Examples of sliding scales with or without sharing. The X-factor is adjusted based on measured rate-of-return (ROR).

Optionally, the regulator can apply a sharing mechanism where the X-factor is adjusted only partially in the case that profits exceed the predefined band. In that case, the firm would be allowed to keep a part of the profits achieved in excess of the maximum level. Conversely, in case that the firm earns less than the minimum profit, it would be forced to absorb part of the losses.

The sliding scale strategy assures that profits remain within certain limits but also has the problem that it does not provide any (strong) incentives for the firm to perform in excess of these limits. The firm will not pursue any further productivity improvements once the maximum profit has been attained. In the case that sharing is applied, the firm only has limited incentives as it keeps only a fraction of the realised improvements. From the firm's point of view, additional improvements come at higher efforts but are not necessarily associated with any rewards. Similarly, the firm may well opt for the guaranteed minimum profit level (if this level if sufficiently high) rather than investing in productivity improvement. These problems become particular relevant in the case that the maximum and minimum of the profit range are set too low and high respectively.[10]

3.3.6 Evaluation

The four price-cap strategies differ in the strength of the efficiency incentives they provide. Generally speaking, efficiency incentives are stronger if the relation between price (or X-factor) and costs is more loose and if competitive pressure is increased. In this respect, yardstick competition provides the strongest incentives. Here, the individual firm's price has no relation to its own costs but rather depends on the average of other firms. Also, direct competitive pressure is introduced. However, for the yardstick system to be fair, firms should first be brought to comparable productivity levels. This can be achieved by the related cap system that, in this sense, forms the preparatory phase before entering yardstick competition. Under both yardstick competition and related caps however, the direct link between the efficiency score and the X-factor requires the regulator to be convinced that the benchmarking results are genuine i.e. not driven by modelling or data errors. If this is not the case, it may be more appropriate to adopt a less direct link between the benchmarking results and the X-factor i.e. choose either the isolated caps or sliding scale strategy. Under these strategies, the role of benchmarking is less formal and efficiency scores would provide an indication rather than an exact pinpointing of the productivity improvement potential.

Under the isolated caps approach, rather than directly transferring benchmarking results into the X-factor, the regulator would recognise that benchmarking results are imperfect and use these as the starting rather than ending point for setting the X-factor. Under the sliding scale approach, a formal buffer is imposed between the benchmarking analysis and the X-factor by setting a minimum and maximum allowed level of profit. Based on this, the X-factor would be set at the end of the regulatory period – reflecting the fact that the regulator cannot measure the true performance of the firm. This strategy strongly resembles rate-of-return regulation with the notable difference that the firm can now retain part of its efficiency gains (up to the maximum of the sliding scale). Note that in the extreme case, if the maximum and minimum profit levels were equal, the sliding-scale strategy would coincide with a traditional rate-of-return system.

Sliding scales are the least effective of the four strategies in terms of efficiency properties but score best in the light of financial sustainability and distributional concerns. A sliding scale puts a maximum on the profits of the firm – thus limiting any distributional problems – and also guarantees a minimum profit level – thus in principle guaranteeing a minimum rate-of-return. Isolated caps are more effective in efficiency terms but if the X-factor is not set optimally, this can lead to financial sustainability and distributional problems. Here, there is no limit to the return that the firm could earn and thus the necessity for the regulator to set a proper X-factor is increased. Related caps impose a formal link between the efficiency score and the X-factor and – as a degree of competitive pressure is introduced – provide stronger efficiency incentives. At the same time, there may be serious financial sustainability and distributional problems if this X-factor is set inadequately.

If the benchmarking analysis is incorrect, overestimation or underestimation of the X-factor can cause financial sustainability and distributional problems, respectively. In the former case, the X-factor may be too high for the firm to accomplish and this may cause financial stress. In the latter case, the firm may end up earning windfall profits. Finally, yardstick competition has the most favourable efficiency incentives and is in principle comparable to competition. Its effectiveness depends on the regulatory ability to derive a proper measure of average costs. In that case, there are in principle no distributional problems as the total level of profit would be predefined and included in the measure of average costs. A drawback of yardstick competition, however, is that the continuity of service provision may be at stake, as possible bankruptcy of ill-performing or over-harshly regulated firms cannot be excluded under yardstick competition.

The analysis here suggest, as was also identified in chapter two, that there is a trade-off between incentives and rents. Here, the channels through which this trade-off takes place have been identified. When moving from sliding scales to isolated caps, to related caps, and up to yardstick competition, there is an increase in the efficiency properties of the regulatory strategy but also an increased risks of under or excess profits. Making this trade-off is constrained by information: Better information about the firm's true productivity potential enables the regulator to opt for a price-cap strategy with superior efficiency properties and to extract more rents from the firm. In obtaining this information, benchmarking is an important regulatory asset. The better the benchmarking analysis, the more information can potentially be revealed and the more effective the choice and specification of the regulatory strategy can be.

3.4 Price-Cap Approaches

3.4.1 Calculating the X-factor

The price-cap mechanism aims to provide incentives for better productivity performance. Underlying this objective is the assumption that the firm is actually able to control its level of costs. This may not be necessarily true for all types of costs. There may be some costs that are beyond the firm's control and therefore, it would not be reasonable to expect any productivity improvements in this area. Such non-controllable costs may include items such as taxes, regulatory contributions, fees for connection to other networks (e.g. transmission networks or neighbouring distributors), and costs resulting from so-called force majeure events (e.g. natural disasters).

The incentives would only need to apply to controllable cost items, non-controllable costs would be allowed to be passed through to consumers on the basis of actual costs. The definition of non-controllable costs is not free from ambiguity, however. It may be that some costs are considered non-controllable while in reality, these costs can be (partially) influenced by the firm. Take for example network losses. These costs are driven by two factors namely the amount of physical losses (measured in kWh) and the price paid per unit of kWh of losses. Both these factors are to a certain extent controllable by the firm. For example, operating at higher voltage levels, increasing network capacity, or using better equipment can reduce physical losses.

Although there may be some investments involved in doing so, the fact remains that these losses can (to some extent) be influenced by the firm. Similarly, the price paid for network losses may be reduced by working out better electricity purchase deals to fuel these losses. If the regulator would consider network losses fully non-controllable, the firm would have no incentive to reduce these losses nor to purchase the electricity efficiently i.e. at lowest price possible. Furthermore, some adverse incentives may arise. The firm could for example deliberately purchase losses at a higher price from affiliated electricity selling firms.

With respect to controllable costs, the firm can increase productivity through its own efforts. Controllable costs thus can be regulated on the basis of any of the four price-cap strategies. Generally, regulators distinguish between two types of controllable costs namely costs that are controllable in the short-term (operational expenditures – opex), and costs that are controllable only in the longer term (investments or capital expenditures – capex).[11] Opex typically includes the costs of personnel, maintenance, and overhead costs such as buildings and office rentals, administration, transportation, etc. The firm could adjust its level of opex in a relatively short period. For example, it could immediately reduce its maintenance activities, dispose of personnel, or attract additional staff.

Capex has a long-term nature and is controllable only in the longer run; in the short run, capex can be considered fixed. These costs typically relate to investments for extending network capacity as well as for upgrading quality. Two components of capex can be identified. Firstly, depreciation, which is the return of the investment. The general idea is that, during some predetermined period (the depreciation period) the firm earns back the price that it paid for the investment. In its simplest form, the annual depreciation would be equal to the purchase price of the asset, divided by the depreciation period of that asset. The second capex component is the return on the investment; this is generally defined as an annual rate-of-return on the un-depreciated portion of the investment. The rate-of-return is typically set by the regulator based on an assessment of the firm's costs of capital. Firms have two sources to finance their investments. Firstly, they can attract debt and secondly, they attract equity. For these finance sources, the firm has to pay interest and a dividend respectively. These costs combined determine the firm's costs of capital.[12]

In addition to productivity, both opex and capex levels will be driven by demand. Ceteris paribus, if demand increases, then the absolute level of costs will also increase. At the same time, in the face of scale economies, the increase in costs will be less than proportional to the increase in demand.[13] In the setting of the X-factor, the regulator would in principle need to determine productivity improvement potential as a function of demand growth. There are, however, some problems involved here. If the regulator sets the X-factor at the start of the regulatory period (e.g. as under isolated or related caps), a forecast of future demand will need to be made. Given that such forecasts would be imperfect, there will be an error in the regulator's determination of the X-factor. Forecast errors can be corrected at the end of the regulatory period by redoing the calculation of the X-factor using actual demand figures. Alternatively, the regulator may correct for demand forecast errors in the intermediate i.e. from year to year within the regulatory period.

Although the effects of uncertainty in demand can be corrected for in the longer term, they may cause risks to both firms and consumers in the shorter term. The form of the price-cap scheme largely determines the way in which these risks are allocated between firm and consumers. That is, the way in which differences between forecasted and actual demand are treated. For example, the regulator could choose to completely ignore such differences. This particular form of price-cap is known as the pure price-cap. In this case, assuming economies of scale, the firm would bear all risks associated with forecast errors. If demand growth is lower than expected, the firm's revenues will suffer and its profitability may be jeopardised. Similarly, the firm could earn additional profits in case that actual demand is higher than forecasted.

Typically, regulators tend to mitigate the risks of errors in demand forecast by including so-called revenue drivers in the price-cap formula.[14] Here, actual prices or revenues are adjusted periodically on the basis of factors that are assumed to drive costs such as the number of consumers or energy delivered. Although demand forecast risks and consequently the form of the price-cap are important considerations in the design of price-caps, in the remainder of this thesis, the assumption will be that future demand is perfectly known. The central themes here are the measurement of productivity improvement potentials and the translation of this into the X-factor.

The X-factor can be defined in terms of a gradual change in the price towards a level that corresponds to the efficient level of costs. This suggests a smooth gliding path from the existing towards the efficiently deemed price level. However, consideration needs to be given to the fact that costs levels are driven by demand, which may vary from year to year. Also, given that the price includes non-controllable costs, prices are not likely to follow a smooth course over time but will rather fluctuate from year to year. For reasons of price stability however, such fluctuations are generally not desirable. To achieve a gradual change in prices, the X-factor is typically calculated in such a way that the net present value of the firm's revenues and the allowed costs during the regulatory period is equal to zero. Starting from the initial price p_0, the X-factor is then set as follows:

$$\sum_t \left(\frac{1}{(1+r)^t} \cdot (1-X)^t \cdot p_0 \cdot q_\tau \right) = \sum_t \left(\frac{1}{(1+r)^t} \cdot (opex_t + ror \cdot RAB_t + D_t + others_t \right) \qquad (3\text{-}3)$$

Here, t is the year, r stands for the discount rate in the present value calculations, X is the X-factor, p_0 is the price in the initial year, q_τ is the corresponding volume for year τ.[15] The total revenue in present value terms generated by the firm is given by the left-hand side of the formula and is set equal to the present value of the allowed costs. These costs include opex, a rate-of-return (*ror*) on the Regulatory Asset Base (RAB) and an allowance for depreciation (*D*) as well for other costs that are considered non-controllable. The RAB reflects the net value of the investments undertaken by the firm; it is adjusted annually to take into consideration new investments (*Inv*) as well as depreciation (*Dep*). This can be represented in the following way:[16]

$$RAB_{t+1} = RAB_t + Inv_t - Dep_t \tag{3-4}$$

Simply stated, the essence of the price-cap system comes down to determining an appropriate level for each cost component during each year of the regulatory period. This is then in turn reflected in the X-factor. Given that the parameters p_0, q, and *ror* as well as the initial RAB are known, the X-factor would then simply be calculated by identifying an appropriate level for the annual opex, the RAB, and the allowed depreciations.[17] This effectively boils down to making two decisions: Determining the efficient opex level and determining the efficient level of investments that should be annually allowed to enter the RAB. For this purpose, the regulator could apply any of the four price-cap strategies discussed in the previous sections.

Given that there are two cost categories to be regulated, two basic approaches can be identified.[18] Firstly, the regulator could separately assess opex and capex (investments) using one or a combination of price-cap strategies.[19] Essentially, the price-cap can then be thought to consist of two components or building blocks namely an allowance for opex and an allowance for capex (which would consist of depreciation plus a rate-of-return on the RAB). This approach is generally known as building blocks. The second approach is one where the regulator considers opex and capex in an integrated fashion i.e. does not distinguish between them. Here, the sum of opex and capex would be regulated on the basis of a single price-cap strategy. This approach is known as the total costs or totex approach. The two approaches are now discussed in more detail and then evaluated with respect to their incentives for quality.

3.4.2 Building Blocks Approach

Under the building blocks approach, the regulator needs to assess an efficient level of opex as well as an efficient level of capex. In the determination of the efficient opex, regulators tend to make use of benchmarking analysis but there is typically no formal translation of efficiency scores into efficiency improvement targets; there is often some room for regulatory discretion in translating efficiency scores into the X-factor (Jamasb and Pollit 2000, KEMA 2004). The general approach can therefore be classified as one of isolated caps. The notable exceptions are the Netherlands and Norway where the efficiency scores play a formal role in the determination of the X-factor. In both jurisdictions however, the benchmarking applies to totex and not to opex alone.

For investments, regulators typically set the allowed level on the basis of the firm's own investment projections. That is, capex is treated more or less as a pass-through item. At the start of the regulatory period, the firm would be asked to provide the regulator with an overview of its intended investments during the next regulatory period. The regulator may then develop a view of which investments to include in the RAB or simply accept the firm's projection as it is. Investments that have been allowed into the RAB will be completely recouped through the allowed depreciation while the firm would also earn a rate-of-return over the un-depreciated portion of these investments. Related caps and yardstick competition strategies are, as far as known, not applied to capex. The explanation for this is that there are some important problems attached to the benchmarking of capex. This issue is explored in more detail in chapter six.

Treating capex on the basis of pass-through costs creates adverse incentives for the firm to overstate its investment projections. The more investments are included in the RAB, the higher the capital base of the firm will be and the higher will be the level of return that can be earned. This is an important problem as the firm will be tempted to overstate investments in order to maximise future additions to the RAB and therefore boost profits.[20] A related problem is that the firm may also try to strategically allocate operational related expenditure under capex if the regulatory strategy for the latter cost category is less strict. This removes some of the opex from the incentive regime while it also leads the firm to appear more efficient in opex terms and obtain a higher efficiency score. Furthermore, by capitalising opex, the firm can further inflate its RAB and consequently earn higher returns. Recent surveys by Jamasb et al. (2003) and Jamasb et al. (2004) find that such strategic allocations of opex under capital expenditures have indeed been reported by regulators. Empirical evidence of such behaviour is presented by Burns and Davies (1998) for the case of the UK.

Once the X-factor has been set, the firm would in principle be free to decide its own investment level. At the end of the regulatory period, it may well be that the firm invested less than originally planned. This difference would, as the firm may claim, be due to higher productivity and therefore, in the spirit of price-cap regulation, should be awarded to the firm. Although this line of reasoning is in principle correct, it may be that (part of) the resulting savings are in fact driven by inflated investment projections or deliberate under-investments rather than genuine productivity improvements. To mitigate this problem, an investment target could be imposed. If actual investments turn out to be lower than the target, then prices are accordingly adjusted downwards. Similarly, no ex post allowances would be provided for investments in excess of the target. Alternatively, the regulator could impose a band of desired investment levels with a minimum and maximum target; investments exceeding this band would not or would only be partially allowed into the RAB. This approach resembles the sliding-scale strategy and although this would in principle deal with the problems of over- and under-investment, it comes at the cost of weaker incentives at the capex front. The regulator would (partially) claw back cost savings irrespective whether these are the result of strategic under-investing or due to genuine productivity improvements. This makes it unattractive for the firm to achieve any productivity improvements in the area of capex as there would not be any financial rewards attached to this anyhow.

Building Blocks Approach

Regulatory parameters	
Opex Efficiency Score	90%
Annual reduction in opex	2.6%
Depreciation period	20 years
ror	10%

	Year 0	Year 1	Year 2	Year 3	Year 4
Allowed Investment (I)		**40.0**	**60.0**	**40.0**	**80.0**
Allowed Depreciation (D)					
From previsous investments		50.0	50.0	50.0	50.0
From investments in year 1		2.0	2.0	2.0	2.0
From investments in year 2			3.0	3.0	3.0
From investments in year 3				2.0	2.0
From investments in year 4					4.0
Total Allowed Depreciation		**52.0**	**55.0**	**57.0**	**61.0**
RAB Calculation					
Starting RAB	- / -	1,000	988	993	976
Plus: New investments		40	60	40	80
Minus: Depreciations		52	55	57	61
Ending RAB	1000	988	993	976	995
Average RAB	1000	**994**	**991**	**985**	**986**
Calculation of Building Blocks					
Opex	100	97.4	94.9	92.4	90.0
Depreciation	50	52.0	55.0	57.0	61.0
Returns (ror * RAB)	100	99.4	99.1	98.5	98.6
Allowed Revenues	**250**	**248.8**	**248.9**	**247.9**	**249.6**

Total Cost Approach

Regulatory parameters	
Totex Efficiency Score	85%
Annual reduction in totex	4.0%

	Year 0	Year 1	Year 2	Year 3	Year 4
Opex	100.0	96.0	92.2	88.5	85.0
Depreciation	50.0	48.0	46.1	44.3	42.5
Returns (ror * RAB)	100.0	96.0	92.2	88.5	85.0
Allowed Revenues	**250.0**	**240.0**	**230.5**	**221.3**	**212.5**

Figure 3-5. Simplified example of calculations under building blocks and totex. Under building blocks, the regulator sets separate targets for opex and investments. The firm has to reduce its opex by 10 percent in year 4. Annual allowed levels of investments are 40, 60, 40, and 80 in the years 1 to 4, respectively. Under totex, the regulator treats opex and capex in an integrated fashion. Here, a single efficiency target is applied to the sum of opex and capex. The firm needs to reduce totex levels to 85 percent of the initial level.

Box 3-2. Example of the complexity in assessing investment proposals.

Consider a certain area where due to increased demand, the capacity of two parallel transformers in the main feeding station has to be increased. Presently, as shown in Figure (a), each transformer has a capacity of 20 MVA while peak demand is 15 MVA. In case of malfunction of one of the two transformers, electricity supply would not be interrupted as a single transformer has a capacity higher than the peak load. In addition, in case of maintenance to one transformer, there will be no interruptions. This situation is known as the N-1 condition: No interruptions will take place if one component is taken out of service. Suppose now, as shown in Figure (b), that peak demand is projected to grow to 25 MVA. In that case, failure of one of the transformers will lead to an interruption. Thus, the N-1 condition will not be met; a single transformer will not have sufficient capacity to supply the peak load. Furthermore, the interruption is likely to be lengthy of duration given that the transformer needs to be repaired or replaced by a new transformer in order to restore supply. Anticipating these problems, the distribution firm decides to increase the transformer capacity in order to maintain the N-1 standard. This will be done according to the scheme shown in Figure (c). The two transformers will be replaced by units of higher capacity namely 30 MVA. In this case, the N-1 condition would be met.

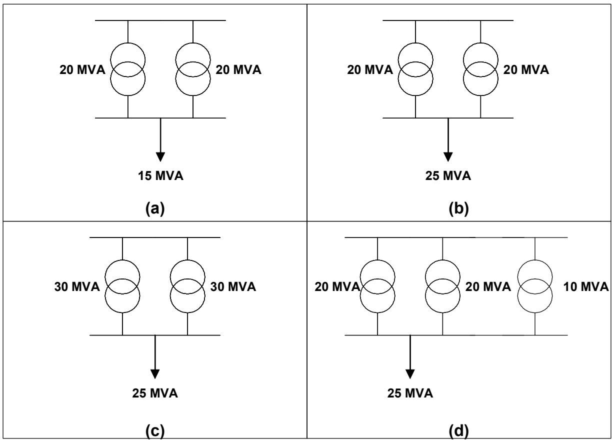


Figure 3-6. Possible configurations for feeding station.

In order to cover the costs of the new transformers, the firm requests the regulator to recognise these costs in the determination of the RAB. The regulator, however, may not consider the firm's solution for the problem the most appropriate one. For example, the regulator could argue that rather than replacing the two transformers, the firm might have chosen to install an additional third transformer of capacity 10 MVA. In this case, which is shown in Figure (d), if anyone of the three transformer fail, total transforming capacity will be at least 20+10=30 MVA. This is a cheaper configuration and also complies with the N-1 condition.

The regulator could also question the necessity of the investment in the first place. He could argue that the firm could well sustain the existing configuration and no investments are needed. Given that the probability of a fault in the transformer is very small, the regulator may take the view that the costs of investing in additional capacity may not outweigh the benefits of high quality. The firm may in turn argue that, according to its own assessments, the investment is well worth its money. The regulator could then decide to allow the firm to invest, but only limited to the installation of an additional transformer of lower rating e.g. 5 MVA. In response, the firm could claim that this would only increase costs in the longer term: If demand grows beyond the expected 25 MVA, then additional investments would be needed anyhow. It would be more efficient to anticipate this by installing some degree of overcapacity.

Which eventual solution would be most economic, that is, provide the best price and quality outcome, would depend strongly on equipment prices, installation costs, as well as consumers' preferences for quality and future demand growth. As may be observed, the capex assessment problem can already become complicated in the case of this simple example. It is not likely that the regulator could assess all possible configurations or discuss each individual investment decision in detail. This would not only lead to duplication of the firm's planning activities but is only impractical given the regulator's limited technical and financial resources.

In order for the isolated caps or sliding scale strategies to be effective, the investment target (or alternatively, the sliding scale band) would need to be set adequately. Ideally, the investment target should reflect the efficient investment level for each individual investment; this would then be allowed into the RAB. However, one can question whether the regulator will have sufficient information to effectively evaluate all proposed investments.

This problem is particularly relevant given that benchmarking of investments is difficult to apply. Investments are, among others, driven by developments in demand and asset replacement considerations. These factors can be different from firm to firm and tend to distort a comparative analysis. Furthermore, investments are typically lumpy while at the same time, firms usually have some degree of flexibility in timing their investments. The firm may decide to postpone investments or bring them forward in time. If investments between firms would be compared i.e. benchmarked, there is a risk that the analysis would be hampered by such factors. Assessment of the desired investment level thus is a difficult issue – in particular when the quality dimension is taken into account. The example provided in Box 3-2 demonstrates this point.

3.4.3 Totex Approach

Under the totex approach, the regulator does not differentiate between opex and capex anymore but sets the X-factor on the basis on the sum of these i.e. on the basis of total costs (totex). In practical terms, this means that the regulator does not need to consider investment projections by the firm but instead performs a benchmarking analysis of actually incurred levels of totex. The resulting efficiency scores then form the basis for setting future allowed totex levels. The efficiency incentives of the totex approach come from the fact that each regulatory period, the X-factor is set on the basis of performance achieved in previous years. If the firm manages to increase productivity, its efficiency score will be higher in future periods and consequently its X-factor will be lower. This is an important difference from the building blocks approach where problems of assessing capex projections hinder the determination of efficient levels of capex.

Under the totex approach, the problem of investment assessments is effectively bypassed. Furthermore, as the totex approach does not distinguish between opex and capex, the firm (as well as society) may also achieve efficiency gains by trading-off better between labour and capital inputs. In terms of the example presented in Box 3-1, the regulator would simply let the decision whether to opt for a replacement of the transformer or installation of an additional transformer (or any other solution for that matter) solely up to the firm. Under the totex approach, the regulator does not need to develop a view whether a given investment proposal should be allowed or not. Rather, the regulator considers the actual total costs (including investments) incurred by the firm and sets the X-factor based on an analysis of these costs.

Although seemingly attractive, there is an important problem that would need to be considered under the totex approach. This is related to the long-term nature of investments. Capex (depreciation and returns) are spread over a number of years and therefore, the benchmarking analysis would need to consider a long enough period of time rather than taking only a snapshot of costs during a single year. As the example from Table 3-2 shows, the costs in a given year can be strongly influenced by the timing of investments. In this simple example, firms A1 and B1 both invest an amount of 400. In the long run, both firms will face the same level of depreciation costs. However, the firms differ in their timing of the investments. Firm A1 invests primarily in the last year while for firm B1, most investments are conducted in the first year. As can be observed, the effect of this is that firm B1 has high depreciation costs in the early years and relatively low depreciation costs in the later years.[21] If the benchmarking analysis would consider only a single year, say the second year, firm A1 would turn out to be very efficient as it would have much lower costs (50) compared to firm B1 (with costs of 88). The reverse would apply if the benchmarking analysis was conducted later in time e.g. in the last year.

Table 3-2. Simplified example of the impact of different depreciation policies and investment timing. All firms invest the same amount over a period of three years and use straight-line depreciation but differ in the timing of these investments and the choice of depreciation period. Although in the long run depreciation costs are the same, annual depreciation varies considerably.

Firm A1 (depreciates in 4 years)					
		Depreciation (mln. EUR) for investments in year:			Depreciation Costs (mln. EUR)
Year	Investment (mln. EUR)	1	2	3	
1	100	25			25
2	100	25	25		50
3	200	25	25	50	100
4		25	25	50	100
5			25	50	75
6				50	50
Total	400	100	100	200	400

Firm B1 (depreciates in 4 years)					
		Depreciation (mln. EUR) for investments in year:			Depreciation Costs (mln. EUR)
Year	Investment (mln. EUR)	1	2	3	
1	300	75			75
2	50	75	13		88
3	50	75	13	13	100
4		75	13	13	100
5			13	13	25
6				13	13
Total	400	300	50	50	400

Firm A2 (depreciates in 2 years)					
		Depreciation (mln. EUR) for investments in year:			Depreciation Costs (mln. EUR)
Year	Investment (mln. EUR)	1	2	3	
1	100	50			50
2	100	50	50		100
3	200		50	100	150
4				100	100
5					-
6					-
Total	400	100	100	200	400

Firm B2 (depreciates in 2 years)					
		Depreciation (mln. EUR) for investments in year:			Depreciation Costs (mln. EUR)
Year	Investment (mln. EUR)	1	2	3	
1	300	150			150
2	50	150	25		175
3	50		25	25	50
4				25	25
5					-
6					-
Total	400	300	50	50	400

This example demonstrates the importance of considering multiple years in a totex benchmarking analysis. Multi-year analysis, however, makes the benchmarking analysis more data demanding and therefore less practical. Consideration should also be given to the fact that the analysis may be hampered as a result of different accounting conventions in the treatment of capital costs. Consider the above example once more.

Firms A1 and A2 (or B1 and B2) both have the same investment pattern, but use different depreciation periods. Because of this, their depreciation cost measured in the same year tends to be different. Firm A1, which uses a depreciation period of four years, has lower costs in the earlier years than firm A2 that uses a shorter period of two years. Conversely, in the later years, firm A1 still incurs depreciation costs while firm A2 has already depreciated all assets.

Although the examples provided here are very simplified, they illustrate the basic problem of how ignoring the long-term nature of investments can distort the benchmarking results. Even though in the long run firms invest at similar levels, their costs would fluctuate from year to year, reflecting differences in investment timing and accounting policies. Including multiple years in the analysis could solve this issue, but would also make the analysis more data demanding and therefore less practical. This is particularly true if the firms considered in the analysis used different accounting conventions. Performing a backward calculation of book and depreciation values could eliminate monetary effects resulting from such differences. However, there may be a problem in obtaining such historical data, in particular given the relatively long lifetimes of assets in the electricity distribution business.[22]

3.4.4 Integrating Quality into the Price-Cap

Both the building blocks and totex approach aim to provide incentives for higher productivity. The methods to do this are different – the building blocks approach uses a prospective approach and imposes a predefined level of investment for the firm. Investment proposals are evaluated and this leads to the identification of a desired investment level. Under totex on the other hand, the firm is free to decide its own level of investment and future totex targets are set on the basis of an assessment of actually incurred totex.

So far, the quality incentives of the two approaches have not yet been considered. In order to evaluate these incentives, it is helpful to make the distinction between opex and capex. For opex, apart from potential strategic allocation between opex and capex, incentives are in principle the same. Under both approaches, opex targets are set on the basis of an assessment of actual opex levels. It is likely that the stronger the incentives for higher productivity, the higher are the risks of an adverse decrease in quality. For example, one may expect quality problems to be more severe under yardstick competition than under a sliding scale strategy. Under the former, the presence of competitive pressure provides strong incentives to cut costs. Under sliding scale on the other hand, such incentives are less strong as cost reduction at some point would no longer generate any additional profits. Due to its short-term nature, reductions in opex can have immediate effects on quality. Take the example of saving on the costs of repair crews. As there will be less human resources available to restore power in case of interruptions, the average duration of interruptions is likely to increase. Such effects will be noticed immediately. For other operational expenditures however, cost reductions may not be noticed that promptly. Poor maintenance, for example, is very likely to impact the outage performance of network components as this will affect their quality. However, these effects may not be noticed immediately. It generally takes some time before the effects start to become visible.[23]

With respect to capex, the impact on quality is different under building blocks and totex. Under totex, the firm is free to decide its own level of investment. If no quality provisions were made, the logical incentive would be to reduce investments, as this will bring down costs and consequently generates a higher efficiency score in future periods. This is likely to have an adverse impact on quality levels. The firm may not invest sufficiently in capacity expansion of the network. In particular, in times of high demand growth, this may lead to inadequate network capacity levels. Given the typical long lead times of investments, it may take some time to set in place new capacity. In the meantime, however, consumers will experience frequent blackouts in particular during peak hours when the existing capacity is not sufficient to cover all demand. On the other side, the totex approach also provides incentives to cut on replacements and investments in quality upgrade. As a result, the general quality level of network components may decrease and the probability of breakdown will be higher. Combined with reduced maintenance, this may result into serious quality problems that, however, are likely to be noticed only in the longer term.

The source of the quality problem under totex is the fact that the regulator does not include quality in the measurement of productivity. That is, the benchmarking analysis only considers costs and not quality. The basic idea of a price-cap is to provide incentives for higher productivity. If, however, quality were not included in the measurement of productivity, then efficiency scores would not be affected by substandard quality performance. The logical solution then is to include quality into the benchmarking analysis i.e. include quality in the measure of productivity. Then, the firm could still be left free to decide on its spending levels, but in future regulatory periods the impact of these decisions on both costs and quality would have impact on the level of its X-factor. Anticipating the fact that productivity measurement now also incorporates quality, the firm would spend sufficiently on quality in order to drive up its efficiency score.

Investment and quality problems under building blocks are of a different nature. In principle, there is no risk of quality degradation here as the regulator dictates the desired level of investments. An important condition however is that the investment target reflects an appropriate level of quality. In the face of uncertainty whether a given investment leads to a sufficiently high quality level, the regulator may be tempted to increase the investment target. This however leads to a waste in the productivity sense as the firm will be overspending on capex and likely provide a too high level of quality. Determining the optimal investment level is therefore an important condition for assuring both an optimal quality level as an efficient level of investment under the building blocks approach. This evidently requires information about the cost-quality relation as well as knowledge of what the optimal quality level should be. In turn, this requires the ability to simultaneously assess the cost and quality performance of the firm's proposed investments.

3.5 Conclusions

3.5.1 Synthesis

This chapter has developed a taxonomy of price-cap approaches and evaluated these in the light of their quality incentives. Quality problems would arise if, in the setting of the X-factor, the regulator would not include the potential impact of the firm's cost choices on quality. This problem can be solved by integrating quality into the price-cap. As has been shown, the efficiency and quality properties under the building blocks and totex approaches tend to be different. Integration of quality into the price-cap is therefore likely to come in different forms.

For the totex approach, quality integration implies the application of an integrated price-quality benchmarking analysis of the firm's actually incurred costs and quality levels. The benchmarking analysis would somehow need to model quality into the determination of the productivity frontier and consequently in the derivation of the efficiency score. For building blocks, integration comes in the form of conducting an integrated price-quality assessment of the firm's proposed investments. Here, the regulator would need to simultaneously assess whether the implied investment leads to a desirable level of quality, and whether the investment is undertaken at least costs. The heterogeneous nature of investments limits the scope to perform benchmarking analysis across firms as investments are typically very different and difficult to compare against each other. Therefore, the regulator will rather need to assess the cost and quality performance of the firm's investment proposals on an individual basis.

It seems that the quality problem is more severe under totex than under building blocks. Under the former approach, the regulator leaves the discretion of deciding on investments with the firm and subsequently cannot influence the quality level that the firm is providing. If no additional quality provisions exist, the firm's quality choice is likely to be driven by cost reduction considerations rather than by a concern with providing high quality to consumers. Under building blocks on the other hand, the regulator can indirectly steer quality levels by prescribing the desired investment level. However, as determining the desired investment target is inherently difficult, this may well lead to some degree of over-investment. This observation provides insight in the trade-off between price and quality. Under building blocks on the one side, quality degradation can be avoided by allowing a high investment level. This leads to higher prices and is essentially the premium for assuring high quality. Thus, the trade-off is biased towards quality rather than price. Under totex on the other side, the regulator cannot directly control the firm's investment level and therefore not influence quality. At the same time, efficiency incentives are higher as the firm has an incentive to reduce costs to a level that it considers suitable. Thus, under totex, the price-quality trade-off is somewhat biased towards price.

In both cases, the ability to integrate quality into the price-cap can help in developing a more effective regulatory approach. Under totex, incorporating quality into the analysis of productivity allows the regulator to derive an efficiency score that reflects an optimal trade-off between price and quality. Similarly, under building blocks, considering both the price and quality performance of an investment can help to determine an optimal level of the RAB. In the process of integrating price and quality, benchmarking can play an important role. Developing integrated benchmarking tools can assist the regulator in simultaneously evaluating price and quality and in this way, obtain better information to establish a more balanced regulatory approach.

3.5.2 The Way Forward

In order to develop an integrated price-quality benchmarking model for both the totex and building blocks approach, a better understanding of the concept of optimal quality is important. This is the central theme of the next chapter. The results obtained are then applied in chapters five and six where integrated price-quality benchmarking models are developed for the totex and building block approach, respectively.

Notes

[1] Conditions for perfect competition generally do not hold completely. However, the perfect competition case performs a useful benchmark and helps to understand the mechanics at work under competition. See Cabral (2000, pp. 85-86) and Stiglitz and Walsh (2002, pp. 26-27).

[2] Price-caps primarily aim at inducing higher productive efficiency. Allocative efficiency can be achieved through better pricing structures. The latter aspect is not considered here.

[3] In principle, prices are expected to decrease over time i.e. the X-factor is positive. However, in some cases the X-factor can be negative i.e. the price-cap results in a price increase. This may be the case if initially, prices were not at cost-reflective levels. Examples of negative X-factors can be found in some of the former communist states in Eastern Europe (KEMA 2004).

[4] There can be a problem if the inflation index also includes in the good basket the price for electricity. This creates circularity between the electricity price and the inflation index. Therefore, in principle, the inflation index should be adjusted to exclude electricity from the basket of goods.

[5] In the Netherlands for example, the Dutch regulator estimated an average efficiency improvement potential of around five percent while the estimation for the frontier shift was only two percent (DTE 2002a).

[6] This does not necessarily mean that the peers in the given benchmarking sample do not have any scope for further improvement. It may well be that there are other even more efficient firms which were not included in the sample. Furthermore, there is also the frontier shift that needs to be taken into account.

[7] If the regulator needs to set prices at the start of the regulatory period, initially an estimation of the X-factor can be made. At the end of the regulatory period, the X-factor can be adjusted based on realised cost developments. This approach is for example followed in the Netherlands (DTE 2004a).

[8] As far as known, yardstick competition for electricity distribution has only been adopted in the Netherlands. In sectors such as hospitals and universities, it is applied more frequently (CPB 2000).

[9] If there is only a single firm to be regulated, an international benchmarking sample could be used. Also, the benchmarking analysis can for example be applied to the firm's regional branches.

[10] In principle, the isolated cap strategy can be considered a sliding scale without any minimum or maximum i.e. linear throughout the whole range of the firm's profits.

[11] The distinction between short-term and long-term costs, generally denoted as labour and capital, is the usual one in economic theory. See for example Douma and Schreuder (1998) p. 25.

[12] This is usually expressed in terms of the Weighted Average Costs of Capital (WACC) which is the weighted average of debt and equity costs Cost of debt usually follows from an analysis of market interest rates. Cost of equity are more difficult to measure and typically involve the use of the so-called Capital Asset Pricing Model (CAPM). See also Brealy and Meyers (2000) pp. 195-203.

[13] For opex, the relation between demand and costs is more continuous than for capex. Investments are usually lumpy, this makes the demand-capex relation a shock-wise one rather than a continuous one (Turvey 2001).

[14] For a comprehensive overview and discussion of different price-cap forms, see for example Green and Rodriguez-Pardina (1999) or KEMA (2004).

[15] The variables t and τ do not necessarily refer to the same year. The choice of τ in relation to t is an aspect of the form of the PCR control and determines the way risks resulting from volume forecast errors are allocated between firms and consumers. As mentioned earlier, this aspect will not be considered further.

[16] In practice the RAB would also need to incorporate disposals as well as capital contributions. Furthermore, some regulators also include an allowance for the costs of working capital in the definition of the RAB.

[17] The initial price p_0 would have been previously set by the regulator. Demand forecasts would be reflected in q while for the discount rate r, typically the allowed rate-of-return (*ror*) would be used.

[18] In reality, the regulatory would also need to incorporate non-controllable costs into the X-factor. These costs would be treated on the basis of actual costs and therefore not be exposed to any incentives.

[19] The regulator could choose to apply the same price-cap strategy to both opex and capex. However, the assessment of each cost component would still be performed separately.

[20] The RAB and the firm's book value are in principle not the same. The RAB reflects the investments that the regulator considers to be appropriate but these are not necessarily undertaken by the firm.

[21] For simplicity, only depreciation costs are compared. Similar effects would also apply to returns.

[22] In the Netherlands for example, lack of suitable data made accurate corrections of capital costs difficult. This was an important factor for firms to reject the benchmarking analysis conducted by the Dutch regulator (Ajodhia et al. 2003).

[23] Less maintenance may actually increase quality in the short-term as less network components will be taken out of service (KEMA 2002).

4

Quality Regulation

4.1 Introduction[1]

4.1.1 Background

Under an integrated approach, the regulator would set the X-factor directly as a function of quality. The previous chapter presented two possibilities for integrating quality into the X-factor. The first one, which is related to the totex approach, would consist of performing an integrated cost and quality benchmarking analysis of the firm's previous performance. The second possibility, which is related to the building block approach, would consist of integrated cost and quality assessments of the firm's investment proposals. In practice, such fully integrated price-quality approaches have – as far as known – not yet been applied. In contrast, regulators tend to set the X-factor only on the basis of an assessment of the firm's costs and on top of this, apply separate quality controls that aim to drive quality into desirable directions. The use of such quality controls under the price-cap can be considered as a partially integrated price-quality system.

In the context of applying integrated price-quality regulation, in line with Spence (1975), three informational problems play an important role. Firstly, there is the problem of measuring quality. Clearly, if the regulator could not measure quality, it would not be possible to perform an integrated cost and quality analysis. The second problem is that of measuring the relation between cost and quality. Generally speaking, higher costs (e.g. more investments) will produce higher quality levels. However, quantifying this relation is complex as it may differ as a function of the location of the network and change over time. These spatial and temporal variations would need to be taken into account in the development of the price-cap scheme. Furthermore, quality costs would also depend on the output level of the firm as well as on the firm's productivity level. The third informational problem is that of measuring consumer demand for quality. Investments in quality would only be economic if this creates a net benefit to society i.e. consumer willingness to pay for quality improvement is larger than the costs of realising these improvements. To identify whether this is the case and what quality level should be aimed at, information about consumer demand for quality is needed.

The objective of this chapter is to explore the two latter informational problems. A better understanding of these problems can then help in establishing an effective integrated price and quality regulation approach. The quality measurement problem is, relative to the two latter ones, of less immediate interest here and is further explored in Annex I.

4.1.2 Chapter Outline

Section two starts with an assessment of quality controls. Here, an overview of the different controls is provided and these are evaluated in the light of their effectiveness i.e. the extent to which they provide incentives for optimal quality. Some practical experiences in the area of quality controls are discussed in the Annexes. Section three investigates the concept of optimal quality in more detail. Here, the problem of measuring consumer demand for quality is an important aspect. Techniques to measure quality demand are also studied in this section. Section four is mainly related to the problem of measuring the cost and quality relationship. In the context of quality controls, this comes down to the setting of the quality target; different approaches are discussed here. Also, the problems of spatial and temporal variations in the cost-quality relation are assessed.

4.2 Quality Regulation Controls

Overviews of quality controls are contained in Rovizzi and Thompson (1995), Arblaster (1999), IPART (2001), Williamson (2001), DTE (2002b) and Raza (2003). Generally speaking, two classes of quality controls can be distinguished. Firstly, indirect quality controls aim to provide consumers with information about the firm’s quality performance and create institutions through which these better-informed consumers can demand or pressurise the firm to deliver an appropriate quality level. The second class of quality controls are direct controls. Here, the regulator provides the firm with direct financial incentives (penalties or rewards) in order to provide an appropriate quality level. Such direct controls come in the form of minimum standards or incentive schemes.

The main difference between the two classes of quality controls relates to the role of the regulator. Under direct controls, the regulator plays an active role; he develops a view of what quality levels to aim at and provides the firm with incentives to reach these. In contrast, under indirect controls, the role of the regulator is primarily one of an information provider and facilitator of disagreements on quality between firms and consumers. The indirect and direct quality controls are now discussed in more detail.

Table 4-1. Overview of quality regulation instruments in different regulatory jurisdiction. Sources: ACCC (2003), CEER (2003), and personal communications with regulators.

Jurisdiction	Performance publication	Overall Standard	Individual Standard	Quality Incentive Scheme
Australia (Victoria)	Yes	No	Yes	Yes
Belize	Yes	No	No	Yes
England & Wales	Yes	Yes	Yes	Yes
Italy	Yes	Indicative	No	Yes
The Netherlands	Yes	Indicative	Yes	Yes
Norway	Yes	No	No	Yes
Portugal	Yes	Indicative	Yes	No
Spain	Yes	Indicative	Yes	No

4.2.1 Indirect Quality Controls

A widely applied type of indirect quality control is performance publication. Here, the regulator requires the firm to disclose information about (trends in) its quality performance to the public. Overviews of the firm's quality performance are then provided, for example, in the firm's annual reports, in dedicated regulatory publications, or on the firm's or regulator's website. Additionally, the regulator can oblige the firm to take into consideration the views of consumer representation groups or include consumers in the advisory or supervisory boards of the firm. Alternatively, consumers can be provided increased possibilities to express their quality concerns by establishing complaint handling bodies (e.g. consumer hotlines) or institutions where conflicts between consumers and firms can be handled (e.g. an Ombudsman). The complaint handling function may also be carried out by the regulator himself, as is for example the case in Italy.

Indirect controls are relatively simple to implement and do not require the regulator to develop a view on adequate quality standards. The basic idea is to expose the firm to public scrutiny by providing consumers with information about the firm's performance. The assumption is that the firm, considering its reputation or public image, would then be inclined to match its quality to consumer demand. However, this assumption may not always be true. Being a monopolist, unsatisfied consumers would not have much financial impact on the firm; consumers have no alternative and cannot leave the market anyway.[2] The incentive for better performance would therefore not be financial but would have to come from a reputation concern. If such a reputation concern were absent, indirect quality regulation would not be effective.

In the case that the regulated firm is also operating in other competitive businesses, then Weisman (2002) shows that a concern with quality in these competitive markets can spill over to the monopoly market. Consumers served by the monopoly network may also procure services in other competitive markets where the firm also operates. For example, the distribution firm typically has associated firms that are active in the supply or generation business, sometimes even operating under the same (brand) name. In this case, if consumers are dissatisfied with the services by the monopoly network provider, they may be less inclined to procure supply services from these associated firms. On the other hand, if the firm performs well in the monopoly business, this positive reputation would spill over to the competitive markets and create a competitive advantage there.[3]

But even if the firm is concerned about its reputation, it may choose to deal with this in ways that are less costly than increasing quality. The firm could for example choose to influence public opinion through advertising or sponsoring activities. In the extreme case, the firm could even abuse increased public concern with low quality at its own advantage; it could put the blame on a too harsh price-cap system and claim that low quality is caused by the lack of funds provided under the price-cap. Thus, public exposure may fire-back and rather than exposing the firm's performance, question the ability of the regulator.[4]

One may also question whether consumers are capable to comprehend the performance statistics published by the firm. The firm could provide technical explanations – too complex to understand by consumers – to justify its low performance. Consumers are generally not in a position to properly assess the reasonableness of such explanations. The firm may use this informational asymmetry at its own advantage. This problem may be partially overcome if there are multiple firms to be regulated. The regulator could then publish comparisons between firms in order to introduce some degree of competitive pressure. Apart from the assumption of the presence of a reputation concern, one should take into account that part of the performance differences may be explained by differences in operating conditions. Simple comparative statistics without taking into account such differences, would then lead to unfair comparisons.

4.2.2 Minimum Standards

Minimum standards dictate a minimum level to be achieved for a certain performance aspect. In case of not meeting this standard, the firm is financially penalised.[5] In some cases, the standard is only indicative and substandard performance does not lead to any penalty. These so-called indicative standards can, however, be considered as an indirect quality control rather than a minimum standard. Although the regulator would develop a view of desired (minimum) performance levels, enforcement of the standard would effectively be left to consumers.

Minimum standards differ in the way they incorporate the measurement of performance into the standard. Overall standards relate to performance at the aggregate level and would typically be measured using an indicator such as SAIFI or SAIDI.[6] Once performance falls below the standard, there will be a penalty in the form of a price reduction or rebate for all consumers. Overall standards thus measure at the system level; they may not reflect any deviations in quality levels amongst different consumers. This problem can be partially solved by applying the standard at a less aggregated level. For example, standards can be defined per region or by differentiating between rural and urban areas. Usually, areas with higher consumer density (e.g. urban areas) have a higher standard, reflecting the higher costs involved in supplying consumers living in rural and less densely populated areas. Consequently, the minimum standard for urban areas would be set higher than for rural ones.

In the extreme case, the standard could be applied at the level of the individual consumer. This individual standard can then be defined in terms of a limit to the number of interruptions or the duration of these interruptions experienced by an individual consumer. In the case of violation of the standards, penalties would be paid directly to the affected consumers e.g. in the form of a discount on the electricity bill. Although, in financial terms, the effect of a penalty under an overall or an individual standard would be the same to the firm, individual consumer compensation is generally preferred. Here, only consumers who actually experienced substandard performance would be financially compensated. Also, an individual compensation may be more significant to an individual consumer as opposed to a general price decrease where the compensation would be socialised amongst all consumers. These individual compensations can be either automatic or discretionary. In the former case, the firm would register which consumers suffered from substandard performance and automatically pay these consumers compensation. In the latter case, only consumers who claimed would be allowed the compensation. The former approach is generally preferred as consumers may not always be fully aware of the existence of the standards or may find it costly to claim compensation.

An important issue to take into consideration when applying minimum standards is the level of the penalty. If the penalty is too low, the standard is not credible and the firm may find it profitable to ignore the standard altogether. For example, if a minimum standard would set a maximum duration of an interruption to eight hours, the firm would pay the same penalty if the interruption lasted eight hours or eight days. The firm could choose to ignore the minimum standard if the associated cost savings would exceed the penalty. This problem would be further aggravated if the standard were also set too low. Conversely, if the penalty it set too high, this would impose unrealistic targets for the firm and possibly have an adverse financial impact. Too high standards are likely to be costly for the firm both in the sense of meeting them as well as having to pay their associated penalties. Particularly when the penalty level is high, too stringent standards may lead to financial stress and jeopardise the firm's viability.

Minimum standards define clearly the boundaries for quality levels; both firms and consumers know exactly what minimum level of quality to expect. The presence of a penalty can provide firms with strong incentives to deliver adequate quality levels. Still, minimum standards do have a number of limitations and problems. A minimum standard imposes a discrete relation between performance and price. Either the firm pays a fine or not; there is nothing in between. This may have the effect that firms, who have an interest to save costs, may choose to perform at a level close to the minimum standard. That is, they will not supply higher quality levels than strictly required by the standard (possibly taking into account some safety margin). Effectively then, the minimum standard would also implicitly prescribe at which level consumers would be served.

4.2.3 Quality Incentive Schemes

A quality incentive scheme can be considered as an extended minimum standard. Here, a more continuous relation is imposed between price and quality. Each performance level results in a financial incentive, which varies with the gap between actual performance level and some predefined target level. In case the firm performs below the target, the incentive is a financial penalty while if the firm exceeds the target, the incentive comes in the form of a financial reward.

Different types of quality incentive schemes exist. Price and quality can be mapped continuously or in a discrete fashion, the level of the penalty or reward can be capped, dead bands may be applied, etc. Some examples are provided in Figure 4-1.

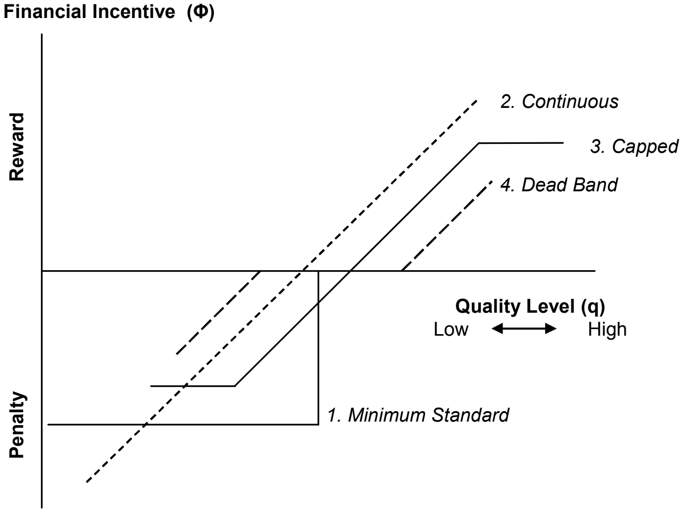


Figure 4-1. Examples of quality incentive schemes. The horizontal axis represents the actual quality performance, the vertical axis the financial incentive.

Under the first scheme, after reaching a certain quality level, a fixed penalty is imposed. This is essentially an ordinary minimum standard. In the second example – the continuous scheme – there is a continuous relation between price and quality. For each level of quality, there is a corresponding penalty or reward which is proportional to the gap between actual and target performance. The third scheme is similar to the second but now with a cap on the level of penalty and reward. Essentially then, the scheme is only linear within a predefined band; outside this band, the scheme is similar to a minimum standard and has similar problems. If quality decreases beyond some minimum level, the penalty paid by the firm does not increase further. Similarly, quality levels exceeding the maximum level would not generate any additional rewards to the firm. Thus, there will be no incentive for the firm to further improve quality once the maximum reward has been reached - even if it would be economic to do so from a social perspective. The fourth scheme has a dead band; quality variations within this band do not lead to price adjustments. The reason for this is to prevent shocks in the level of the financial incentive due to stochastic variations in quality. Stochastic effects can lead to quality fluctuations and consequently also a fluctuation in the level of penalties and rewards. Dead bands can dampen these effects but also create the problem that genuine quality changes – that take place within the dead band – remain undetected. An alternative approach to limit the impact of stochastic quality variations is to use multi-year averages in the quality measurement. This has the same dampening effect as a dead band but also makes sure that all changes in quality are eventually translated into a financial incentive. Alternatively, as is applied in for example Norway, penalties and rewards can be cumulated for a number of years before they are translated in a reduction in prices.

Table 4-2. International experiences with quality incentive schemes for electricity distribution.

	Incentive Scheme	Quality Indicator	Quality Target	Penalty / Reward
Belize	Continuous	SAIFI and SAIDI	Annual improvement of five percent	Penalty around 2.5 EUR / interruption and per minute
Italy	Continuous	SAIDI	Annual improvement target up to 16 percent	Penalty or reward around EUR 18/kWh
Hungary	Discrete with caps	SAIFI and SAIDI	Annual improvement target up to 16 percent	Penalty up to three percent of firm's revenue
Netherlands	Continuous	SAIDI	Average actual performance of the industry	Penalty or reward based on consumer interruption costs
Norway	Continuous	Energy Not Supplied	Historical performance	Penalty or reward 0.40 EUR/kWh (households) and 4.50 Eur/kWh (businesses)
United Kingdom	Continuous with caps	Consumer Interruptions and Minutes Lost	Historical performance	Penalty or reward up to two percent of firm's revenue

There are also different approaches possible for setting the financial incentive. The simplest approach is to fix the incentive per unit of quality performance. For example, if the incentive scheme were based on the number of minutes lost per year (e.g. measured through SAIDI), the incentive would be a fixed amount per minute lost. Alternatively, the incentive can be expressed as a percentage of the revenue or as an increase or reduction in the allowed rate-of-return. Table 4-2 provides the results of a survey of quality incentive schemes used in a number of regulatory jurisdictions. The Annexes contain a number of case studies of quality regulation for electricity distribution.

4.2.4 Optimal Quality Incentive Schemes

Incentive schemes have a clear advantage over minimum standards. As the financial incentive is no longer fixed but made proportional to the difference between actual and targeted performance, there is no incentive anymore to operate just above some minimum quality level. Rather, the firm will seek to provide a quality level based on a trade-off between the costs of providing quality and the resulting penalties or rewards. Thus, through the choice of the financial incentive, the regulator can drive the firm's quality strategy. More specifically, if the incentive level is set at a level that reflects consumer willingness to pay for quality, the firm's quality choice will coincide with what would be desirable from a social point of view. To demonstrate this point, consider the following model.

Let x be the number of consumers; all consumers are provided with a quality level of q. The firm is provided with an incentive Φ, which depends on the difference between its actual performance q and the quality target q_0. For quality levels lower than q_0 the firm pays a penalty ($\Phi<0$), for levels higher than q_0 it receives a reward ($\Phi>0$). The incentive can be defined as a general function of the marginal incentive φ in the following way:

$$\Phi = x \cdot \int_{q_0}^{q} \varphi(q)dq \tag{4-1}$$

The firm's quality choice will be driven by profit maximisation. In line with Equation (2-11), the profit of the firm can then be given by:

$$\pi = x \cdot P - c + \Phi \tag{4-2}$$

Where P is inverse demand and c is the firm's costs. For a given number of consumers, a profit maximising strategy implies choosing a quality level such that:

$$\begin{aligned} &\pi_q = 0 \\ &\Rightarrow x \cdot P_q - c_q + \Phi_q = 0 \end{aligned} \tag{4-3}$$

From the social perspective, the optimal quality level would be achieved if, given a certain output x, total welfare W would be maximised:

$$W_q = \int_0^x P_q dv - c_q = 0 \tag{4-4}$$

The social optimal quality and the firm's private quality choice would coincide in the case that:

$$\begin{aligned} &W_q = \pi_q \\ &\Rightarrow x \cdot P_q - c_q + \Phi_q = \int_0^x P_q dv - c_q \end{aligned} \tag{4-5}$$

This implies that the incentive scheme would need to be configured such that:

$$\Phi_q = \int_0^x P_q dv - x \cdot P_q \tag{4-6}$$

Alternatively:

$$\frac{\Phi_q}{x} = \frac{1}{x} \int_0^x P_q dv - P_q \tag{4-7}$$

This condition implies that the regulator can induce the firm to choose an optimal quality level if the incentive per consumer is defined as the difference between the average and marginal consumer willingness to pay for quality. Effectively then, the incentive transforms all social surplus to the firm i.e. internalises into the firm's decision process the impact of its quality choice on consumer surplus. This approach is conceptually similar to the scheme proposed by Loeb and Magat (1979). Similarly, there would be a distributional problem as consumers would be left with no surplus; all surplus would be allocated to the firm. This problem could be partially solved by setting the term q_0 equal to some baseline quality (e.g. the historical average). In this way, the firm only receives a fraction of the surplus. A more important problem, however, is that of quantifying consumer demand for quality. This, as noted by Spence (1975), is one of the main information problems faced in the application of quality regulation; it is reflected in the problem of estimating the function φ. This issue is further explored in the next section.

4.3 Measuring Demand for Quality

4.3.1 Optimal Network Reliability

In the context of regulating network reliability, measuring quality demand comes down to determining consumer willingness to pay for network reliability. Due to the perceived problems of measuring this, demand for quality is often approximated by its counterpart namely the costs that consumers experience due to interruptions (Munasinghe 1984). Interruption costs are the costs that consumers incur because of reliability being less than perfect. According to Sanghvi (1982), these costs can be divided into two main categories. Firstly, consumers incur short-term interruption costs, either directly or indirectly, as a result of the interruption. These costs can take different forms and are, broadly speaking, either economic or social. Direct impacts are those resulting immediately from the cessation of supply while indirect impacts result from a response to an interruption. Examples of direct economic impacts are lost production, process restart costs, spoilage, etc. Direct social impacts include inconvenience, loss of leisure time, and personal injury or fear. Indirect impacts can be panic, riots, or looting during the blackout. The second category, long-term adaptive response costs, is associated with the changes in consumers' capital stock resulting from measures to mitigate interruption costs. Examples of such actions are installing emergency equipment (e.g. candles or flashlights), protective switchgear, uninterruptible power supplies (UPS) or backup generators.

Using the somewhat simpler (to measure) concept of interruption costs, the optimal incentive for designing the regulatory incentive scheme for a distribution firm can be established in a more practical way. In order to do so, the optimal reliability model developed by Munasinghe (1984) is helpful to consider. This model is based on the idea that higher network reliability reduces interruption costs for consumers but also comes at higher costs. At some quality level, the sum of both interruption costs and network costs – which is defined as total social costs - will be lowest. This is the optimal quality level that one should aim to arrive at. This model is reproduced below.

If the quality (represented by the reliability level *R*) changes, this has two important implications. Firstly, this leads to an increase in the network costs (*NC*), which include the costs of building, operating and maintaining the network system. On the other hand, as *R* rises, the costs experienced by consumers because of interruptions (*IC*) will decrease. An improvement in *R* will also raise consumers' expectations regarding future reliability levels *R** and is also to induce increased electricity demand, which provides additional net benefits of consumption. Changes in *R** may also affect *IC* as consumers adapt their behaviour patterns to reduce interruption costs. Thus, by increasing *R*, it would be possible to trade-off the higher network costs against the decrease in interruption costs and increased net benefits of induced demand.

On the basis of the above assumptions, the following definitions can be developed. Let demand (*D*) for electricity in a given service, network costs (*nc*), and interruption costs (*ic*) be given by:

$$D = D(p, R^*) \tag{4-8}$$

$$nc = nc(D, R\ ) \tag{4-9}$$

$$ic = ic(D, R, R^*) \tag{4-10}$$

Where p is the price per unit of electricity and R^* is the reliability level that consumers expect to receive. The net benefits of electricity consumption (*NB*) can be written as:

$$NB(D,R) = TB(D) - nc(D,R) - ic(D,R,R^*) \tag{4-11}$$

Where *TB* is the total benefit of consumption. Allowing R to vary, the first-order (necessary) condition for maximisation of net benefits is:

$$\frac{dNB}{dR} = 0 \tag{4-12}$$

Which can be re-written as:

$$\frac{dnc}{dR} = -\frac{dic}{dR} + \frac{dTB}{dR} \tag{4-13}$$

Where:

$$\frac{dnc}{dR} = \frac{\partial nc}{\partial R} + \frac{\partial R^*}{\partial R} \cdot \frac{\partial D}{\partial R^*} \cdot \frac{\partial nc}{\partial D} \tag{4-14}$$

$$\frac{dic}{dR} = \frac{\partial ic}{\partial R} + \frac{\partial R^*}{\partial R} \cdot \left(\frac{\partial D}{\partial R^*} \cdot \frac{\partial ic}{\partial D} + \frac{\partial ic}{\partial R^*} \right) \tag{4-15}$$

And:

$$\frac{dTB}{dR} = \frac{\partial R^*}{\partial R} \cdot \frac{\partial D}{\partial R^*} \cdot \frac{\partial TB}{\partial D} \tag{4-16}$$

The term dnc/dR represents the change in network costs due to variations in reliability levels and consists of two components. Firstly, the direct effect of R on *nc* and secondly the indirect effect via the chain of interactions in which R affects the consumers expectation of reliability R^*, which in turn affects the demand, and then finally causes a change in supply costs. The term dic/dR is the change in interruption costs with respect to reliability and has three components. The first two components may be interpreted analogously to the corresponding components of dnc/dR. The last part $\partial R^*/\partial R \cdot \partial ic/\partial R^*$ represents the change in *ic* due to changes in reliability expectation R^*, which are themselves caused by variations in R.

The term dTB/dR denotes the change in total benefits caused by the induced demand changes arising from variations in R^*. It is important to note the difference between *TB* and *ic*. The effect of R on *TB* is based on the fact that consumers will alter their demand level *D* because their reliability expectation R^* has changed. In other words, their demand curve has shifted because their perception of interruption costs has altered, and this change in *TB* reflects the long-term change in expected interruption costs. Therefore the term *ic* in Equation (4-15) should be interpreted as the remaining short-term unavoidable interruption costs, arising from the difference between the actual and expected reliability levels.

The term $\partial R^* / \partial R$ represents the change in reliability expectation due to variations in *R*. For the static case, which is of interest here, this term can assumed to be zero. Assume further that the effects due to induced demand are ignored i.e. $\partial D / \partial R = 0$. Equation (4-13) then becomes:

$$\frac{\partial nc}{\partial R} = -\frac{\partial ic}{\partial R} \tag{4-17}$$

This condition implies that optimal quality is achieved if the additional costs to provide higher quality are equal to the resulting decrease in interruption costs experienced by consumers. That is, it would only make sense to increase quality as long as this leads to a net decrease in total social costs. If quality is higher than the optimum, there is a welfare loss as consumers would be provided a level of quality where the additional costs of providing this high quality exceed the associated reduction in interruption costs. Conversely, if quality is below the optimum, there is also a welfare loss as quality could be increased at a level of costs that is lower than the reduction in interruption costs. This concept can also be represented graphically in Figure 4-2. Assume that the total costs to society *sotex* is defined as *sotex=nc+ic*. In the case that the level of sotex is minimal, total costs to society will be minimised; this also defines the optimal reliability level.

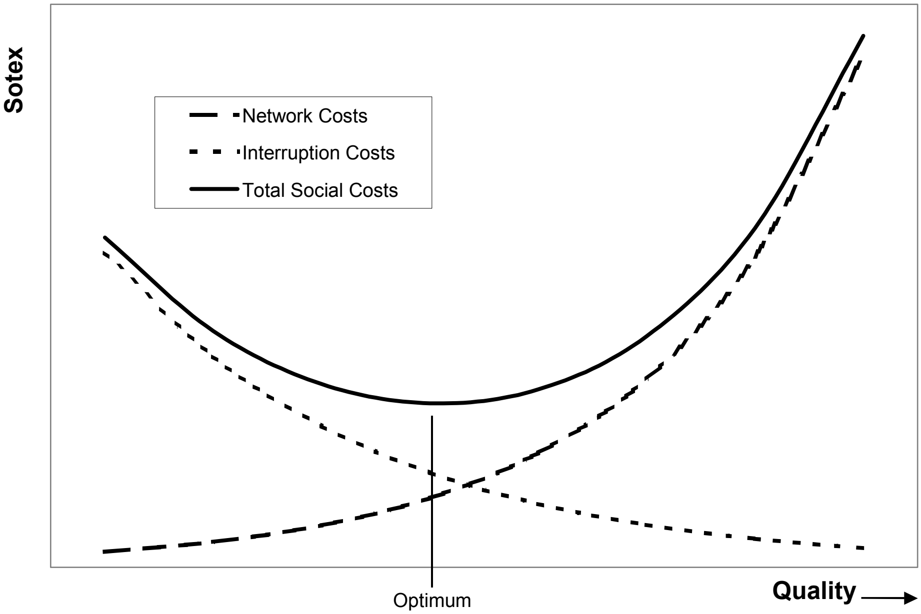


Figure 4-2. Concept of optimal quality.

On the basis of Munasinghe's model, the earlier derived condition for the optimal incentive scheme can be simplified. Profits can be defined as:

$$\pi = p \cdot D - nc + \Phi \tag{4-18}$$

Profits would be maximised if:[7]

$$\begin{aligned} &\pi_R = 0 \\ &\Rightarrow nc_R = \Phi_R \end{aligned} \tag{4-19}$$

And to assure that profit maximisation also leads to a socially desirable quality level, the incentive scheme would be configured as follows:

$$\Phi_R = -ic_R \tag{4-20}$$

That is, the financial incentive should be equal to the total interruption costs that consumers experience at the given reliability level.[8] Effectively, the impact of quality choice on consumer interruption costs would be internalised into the firm's decision-making process. The firm's profit maximisation strategy would then automatically lead it to choose the social optimal reliability level as then, its total profits would be maximised. Furthermore, the regulator would in principle not even need to know the firm's costs as long as he could set the incentive equal to the actual level of interruption costs.[9] Clearly, the effectiveness of this approach will depend on the regulator's ability to properly measure interruption costs and consequently set the optimal incentive level.

4.3.2 Interruption Costs

An interruption takes place when due to a shortage consumers are delivered with less electricity from the power system than originally planned. If the actual consumption is zero then the interruption is full, otherwise the interruption is partial. Sanghvi (1982) distinguished two variables that affect the cost of an interruption namely the type of the shortage and the shortage management strategy.

The shortage can either be capacity or energy related. Capacity shortages relate to situations where the available capacity is lower than peak load. These situations can for example result from generation or network failures or from insufficient investment in network capacity. An energy shortage on the other hand occurs when the amount of electricity that would be purchased during some given period, exceeds the energy available during that period. These shortages are often related to fuel shortages or low reservoir water levels in hydroelectric plants.

From the consumer's perspective, the results of a capacity or energy shortage can take different forms, depending on the shortage management strategy employed by the system e.g. peak shaving, rotating blackouts, interruptions. Different shortage management strategies have a different impact on interruption costs. For example, constraining peak demand – by the price mechanism or by rationing – to a level at which the operating reserve is equal to the normal margin, results in consumers reducing or shifting their peak demand. Alternatively, reducing the operating reserve margin – leading to a situation of unchecked reliability degradation – can lead to more frequent and persistent interruptions with no warning. The costs under the former are likely to be less than in the latter. This is because under the former shortage management strategy any interruption of energy service provision arises on a planned basis whilst in the latter strategy degradation of service reliability entails unexpected interruptions.

In the case of electricity networks, it is primarily capacity shortages that are of concern. Interruptions are caused by faults in network components. The location where the fault occurs drives the impact of the interruption and therefore the level of interruption costs. For example, failures in the transmission network rarely result in interruptions due to the high redundancy. These networks are often operated on the basis of contingency criteria (e.g. N-1), which require that a failure in any random component should not lead to an interruption. Distribution networks are typically operated at lower redundancy levels and consequently have a higher probability of experiencing an interruption due to component faults. However, the resulting interruptions generally affect a smaller number of consumers compared to interruptions at the transmission level.

4.3.3 Measurement Techniques

The literature presents a large number of techniques to measure interruption costs; some common techniques are discussed in this section. A distinction is made between indirect methods and direct or survey methods. Survey methods acquire interruption cost information directly from consumers while indirect methods use other information sources for this purpose. Surveys are again divided into ex post and ex ante surveys, which refer to requesting consumer information about actual and hypothetical interruptions, respectively.

Indirect - Proxies

Proxy methods use indirect data to derive information on interruption costs. In recent decades a couple of proxies have been developed. The ratio of Gross National Product (GNP) to the electricity consumed forms roughly the upper bound for the interruption costs (Shipley et al. 1972, Telson 1975). The ratio of the electricity bill and the energy consumption then provides the lower bound. For residential consumers, the wage rate has been used as a measure of the foregone leisure in case of an interruption (Munasinghe 1980) or the value of lost production for a firm during an interruption (Munasinghe 1981). Loss of production has also been applied to households (Gilmer and Mack 1983).

Indirect - Consumer Surplus Methods

Consumer surplus methods derive interruption costs information from electricity demand curves. The idea is that the willingness-to-pay for electricity depends on the degree to which the consumption of each unit can be deferred to another hour. When elasticity is low, the consumer surplus losses – which are equivalent to the households' willingness-to-pay to avoid a total interruption in that hour – are larger. The consumer surplus losses minus the bill savings provide a measure of the interruption costs (Sanghvi 1982).

Indirect - Costs of Backup Power

Consumers may take preparatory actions to prevent the costs that arise from interruptions by installing backup power. Bental and Ravid (1982) suggest that a profit maximising firm will invest in backup power until the expected gain from the marginal self-generated kWh is also the expected loss of the marginal kWh that is not supplied to that firm. The marginal cost of generating its own power may then serve as an estimate for the marginal interruption costs.

Ex post Surveys - Blackout Studies

Blackout studies collect information about interruption costs from actual interruptions. This method is usually applied in case of large-scale interruptions. Next to quantifying costs, blackout studies often also study the societal impact and preparedness for large interruptions such as police and fire responsiveness, environmental damage etc. Examples of blackout studies include SCI (1978), Steetskamp and Van Wijk (1994) and CEC (1997).

Ex ante Surveys - Direct Costs

Direct cost surveys request interruption costs directly from consumers. Firstly, consumers are requested to identify the different costs categories in case of an interruption. For industrial and commercial consumers these may be lost sales or production, spoilage, damage, etc. The second step is to attach an economic value to each cost category. Interruption costs are then obtained by summing up all the individual costs. Optionally, a list of possible measures and associated costs can be provided and consumers are asked to indicate which measure they would employ for different interruption scenarios.

Ex ante Surveys - Econometric

Two main econometric methods exist. Under the contingency ranking method, consumers are asked to value reliability as if there were a market for it. Thus, a hypothetical market is created where consumers are asked to indicate their willingness-to-pay (WTP) for higher reliability, or willingness-to-accept (WTA) lower reliability levels. Conjoint analysis is similar to contingency valuation with the difference that the WTA and WTP figures are derived indirectly. Here consumers are requested to rank in order of preference different mutually exclusive combinations of price and reliability levels – the price range is determined ex ante by the researcher.

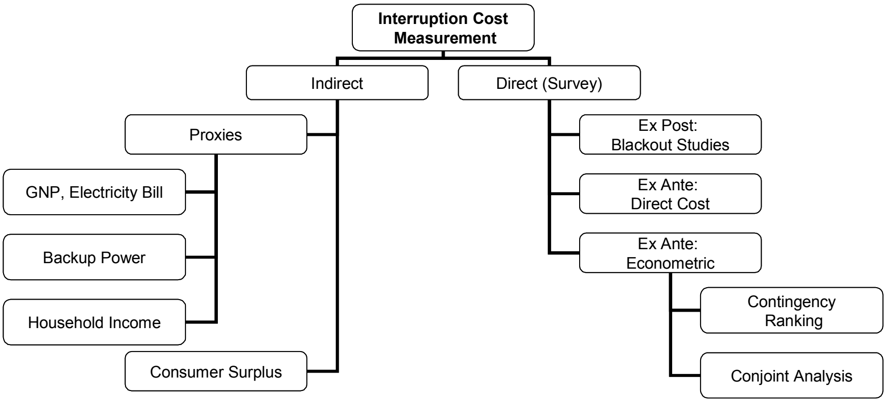


Figure 4-3. Overview of interruption cost measurement techniques.

Evaluation

To compare the different interruption cost measurement techniques, the following criteria can be used: (1) Costs, (2) accuracy of results, and (3) the amount of information that can be acquired. With respect to costs, indirect methods score better. Especially proxy methods require very little and easily obtainable data and thus form an excellent tool to estimate the upper and lower bounds of interruption costs. However, the results are not likely to be very accurate and only give highly aggregated information. Consumer surplus methods require substantial more data than proxy methods although the results may not be proportionally more accurate. There are two fundamental reasons for this: Firstly, the observed willingness-to-pay for planned electricity consumption is not an accurate indicator of what one would be willing-to-pay to avoid an unplanned interruption. Secondly, when measuring system interruption costs, this method assumes that load shedding takes place according to some predetermined order. In practice, this is hardly ever the case (Munasinghe 1981). The cost of backup power method seems to provide a good balance between costs and accuracy. The advantage of this method is that information is revealed from actual consumer behaviour. A disadvantage of this approach is that it is based on the assumptions that firms install generators for backup purposes only and that the installed capacity is below normal peak demand. These assumptions do not always hold in practice. Installed generators often have joint applications while it may well be that installed backup power is equal or higher than peak load due to indivisibility of capacity or low prices of backup power (e.g. UPS technology has improved significantly over the years). Furthermore, this method is primarily relevant for larger consumers as only these are likely to install backup power.

With respect to information, indirect measures score poorly compared to survey methods. Surveys are more expensive to carry out but can deliver quite detailed information about the different factors that influence interruption costs. Blackout studies for example can be used to evaluate the vulnerability of society with regard to an interruption and identify preparatory actions. The problem with blackout studies is that they can only be applied in case of an actual interruption. Ex ante surveys on the other hand, can be planned well ahead in time and can provide substantial amounts of information. The advantage is that the different aspects that impact interruption costs can be studied such as interruption or consumer characteristics.

The main problem of ex ante surveys is their hypothetical character. In theory, the WTP and WTA values that are derived should be the same (Willig 1976). In practice however, it is found that obtained WTP figures are usually equal to zero or otherwise orders-of-magnitudes smaller than WTA figures. Beenstock et al. (1998) argue that the explanation for this can be found in status quo and asymmetry effects. Under the former, the consumer has a resistance to prospective change per se irrespective whether the service is improved or deteriorated. The asymmetry effect (or loss aversion) leads to a difference between WTP and WTA as consumers value prospective service improvements by some fraction of their value of deterioration. This effect can lead to some bias in the survey results.

Table 4-3. Evaluation of different interruption cost methods.

	Costs of the Method	Accuracy of Results	Information Acquired
Proxies	Cheap	Very low	None
Consumer surplus	Costly	Low	None
Blackout	Costly	Reasonable	Reasonable
Direct Costing	Costly	Some bias may exist	High
Ex ante surveys	Costly	Some bias may exist	High

4.3.4 Cost Influence Factors

The costs of an interruption are driven by a number of factors. These factors have been extensively studied in the interruption costs literature – most notably with the use of survey methods. A brief summary now follows.

Duration

As an interruption prolongs, interruption costs increase. Caves et al. (1990) analyse the rate at which these costs increase by comparing results from different studies. For the industrial sector, they find that normalised costs (i.e. per hour of interruption) decrease with duration. This suggests that there is a large initial fixed cost component and a variable component that decreases with duration. Similar comparisons were made for residential, retail, office building, government and farm consumers, which show large inconsistencies between different studies.

Perceived Reliability Level

Another factor that influences the level of costs of an interruption is the reliability level at which the consumer is being supplied. Generally, the higher the reliability level, the more severe the impact of an interruption will be. A study in Nepal showed that 38 percent of residential consumers considered the number of interruptions to be “low” or “very low” although the average number of interruptions was four per week (Pandey and Billinton 1999). Similar results were found in a Brazilian study where more than half of the residential consumers interviewed valued the quality of service provided as “good” although half of these consumers had experienced at least one interruption per month (Gastaldo et al. 2001). In most Western countries, such interruption frequencies would most likely not have produced such high consumer satisfaction ratings. A possible explanation for this is that as the frequency of interruptions increases, consumers can make a better trade-off between expected interruption costs and the adaptive response costs thus minimising total interruption costs. Also, dependency on electricity may not be as high as in Western countries thus leading the relative impact of interruptions to be limited.

Timing

Interruption costs vary with the time of the year, day of the week and time of the day. For residential consumers, winter interruptions lead to higher costs than in the summer while morning or afternoon interruptions are less costly than evening ones (Woo and Pupps 1992). For non-residential consumers, the amount of costs is closely related to the level of firm output. For example, Billinton et al. (1982) find that for retail consumers in Canada the interruption costs during the Christmas season and on Saturdays are significantly higher. An interesting result reported is that for retail and commercial consumers, least costs are incurred during lunchtime (Pandey and Billinton 1999, Gates et al. 1999). For large industrial consumers, the timing of interruptions tends to have little effect; this reflects the constant output delivered in these industries (Dialynas et al. 2001, Gates et al. 1999).

Advance Notice

If an interruption is planned e.g. in case of energy shortages or maintenance activities, advanced notice may be provided to consumers about the occurrence or duration of the interruption. Such actions tend to decrease interruption costs as consumers may take preventive actions or reschedule their original planning. Note that this is in line with the previous observation that consumers experiencing frequent interruptions exhibit lower costs due to increased preparedness. A Scandinavian study reports that planned interruptions can significantly reduce instantaneous interruption costs (Lehtonen and Lehstrom 1995). Similar results have been reported in other countries including the US, Canada and Nepal with reductions varying between 20 and 50 percent (Billinton et al. 1982, Gates et al. 1999, Dialynas et al. 2001).

Consumer Dependency

The degree of consumer dependence on a reliable electricity supply also influences the level of interruption costs. Some consumers may be more dependent than others e.g. hospitals are much more vulnerable for an interruption than a residential consumer. Doane et al. (1988) find – perhaps not unexpectedly – a strong correlation between the presence of electric equipment in a household and the level of interruption costs. Consumers' dependency also increases over time: Sullivan and Sheehan (2000) report a doubling in the real economic quantification of reliability by households in the US over a period of 10 years. Similarly, Andersson and Taylor (1986) report an increase in the real interruption costs from 1969 till 1980 in Sweden.

4.3.5 Cross-Comparison of Interruption Cost Studies

As has been discussed, there are different techniques available to measure interruption costs. Furthermore, the level of these costs tends to vary as a function of different factors. Ideally, these factors should not be considered in isolation as it is more likely a combination of factors that determine the costs that a certain consumer experiences during an interruption. This observation complicates a cross-comparison of interruption cost studies. No interruption is the same; it may differ with respect to its scale, the time it occurs, its duration, etc. Similarly, the type of consumers affected by the interruption will influence the level of costs. These factors may not all be captured (uniformly) by the different interruption cost studies.

Interruption costs themselves may also be presented in different forms. For practical purposes, it is helpful to normalise costs; normalisation can take place in different ways. Some studies normalise interruption costs by the peak load of consumers while others define costs as a function of frequency and duration of the interruption and make a distinction between the fixed and variable costs of the interruption. Most common is to express interruption costs per kWh of non-delivered energy. This approach has also been followed in Table 4-4, which shows the result of a cross-comparison of a number of interruption cost studies.

As may be observed there are substantial variations in the results obtained by different studies. There are different explanations for these large differences. Firstly, the comparison may not be fully compatible due to the fact that numbers had to be converted into a common denominator. Another explanation is the fact that costs may differ by level of economic development, which may differ both by country or region as well as over time. Furthermore, there are differences in the interruption cost studies themselves in terms of the technique that is being used and the scope of the study i.e. the cost-driving factors that have been considered.

The wide variation in results suggests that care should be taken in using the results of interruption cost studies for designing the quality incentive. Ideally, the quality incentive should capture as much as possible the different factors that drive interruption costs and distinguish between different types of consumers. In the theoretical best case, the quality incentive (and therefore the quality scheme) would need to be set for each consumer individually on the basis of the costs incurred by this specific consumer. In practice, however, the regulator should recognise that the quality incentive would need to be set on some average notion of interruption costs. Capturing all possible interruption cost drivers and doing so for each individual consumer is likely to be a too costly undertaking. It would be more practical for the regulator to set the incentive level based on some average measure, possibly differentiated by consumer group. Although such simplifications would possibly distort the incentives, they have the advantage of being relatively simple to apply and easy to comprehend by the firm as well as consumers.

Table 4-4. Cross-comparison of interruption cost studies. All costs are normalised per kWh non-delivered energy and are expressed in 2004 US dollars.[10]

	Methodology	Year	Country	USD / kWh
Residential				
Upadhyay (1996)	Survey	1996	India	0.23
Sarkar and Shreshta (1996)	Survey	1988	India	0.26
Tavanir (1995)	Survey	1995	Iran	2.60
De Nooij et al. (2003)	GDP	2003	The Netherlands	19.35
KEMA (2003)	Survey	2003	The Netherlands	22.99
Young (1987)	Survey	1987	New Zealand	5.25
Turner (1977)	Proxy	1977	New Zealand	1.83
Trengereid (2003)	Survey	2003	Norway	0.48
Shalaan (1989)	Survey	1988	Saudi Arabia	1.29
Andersson and Taylor (1986)	Survey	1980	Sweden	4.18
Lolander (1948)	N/A	1948	Sweden	2.25
Swedish Joint Commission (1969)	Direct	1969	Sweden	4.91
UNIPEDE (1972)	Survey	1970	Sweden	4.30
Sheppard (1967)	Proxy	1965	UK	2.81
UNIPEDE (1972)	Proxy	1970	UK	8.34
Burns and Gross (1990)	Survey	1988	USA	6.70
Krohm (1978)	Black Out	1978	USA	2.88
Faucett et al. (1979)	Black Out	1979	USA	0.13
Sanghvi (1982)	Survey	1980	USA	0.56
Commercial				
Sarkar and Shreshta (1996)	Survey	1988	India	10.12
Tavanir (1995)	Survey	1995	Iran	3.98
De Nooij et al. (2003)	GDP	2003	The Netherlands	9.38
Young (1987)	Survey	1987	New Zealand	31.15

	Methodology	Year	Country	USD / kWh
Trengereid (2003)	Survey	2003	Norway	5.57
Shalaan (1989)	Survey	1991	Saudi Arabia	58.10
Andersson and Taylor (1986)	Survey	1980	Sweden	48.10
Burns and Gross (1990)	Survey	1988	USA	65.67
Industrial				
Sarkar and Shreshta (1996)	Survey	1988	India	9.19
Tavanir (1995)	Survey	1995	Iran	5.25
Young (1987)	Survey	1987	New Zealand	5.25
Turner (1977)	Proxy	1977	New Zealand	5.04
Andersson and Taylor (1986)	Survey	1980	Sweden	18.25
Lolander (1948)	N/A	1948	Sweden	6.48
Swedish Joint Committee (1969)	Survey	1969	Sweden	7.75
UNIPEDE (1972)	Survey	1970	Sweden	10.33
Hsu et al. (1994)	GDP	1991	Taiwan	1.79
Hsu et al. (1994)	Survey	1991	Taiwan	3.37
Taiwan Power Co (1980)	Proxy	1975	Taiwan	1.22
Sheppard (1967)	Proxy	1965	UK	8.38
UNIPEDE (1972)	Proxy	1970	UK	9.99
Jackson and Salvage (1974)	Survey	1970	UK	4.15
Burns and Gross (1990)	Survey	1988	USA	11.22
Grosfeld-Nir and Tishler (1993)	Proxy	1987	USA	17.19
Modern Manufacturing (1969)	Survey	1969	USA	5.80
SCI (1978)	Black Out	1977	USA	13.30
Agricultural				
De Nooij et al. (2003)	GDP	2003	The Netherlands	4.61
Kahn (1997)	Survey	1997	Australia	0.04
Andersson and Taylor (1986)	Survey	1980	Sweden	7.20
Burns and Gross (1990)	Survey	1988	USA	5.84
Whole Economy				
De Nooij et al. (2003)	GDP	2003	The Netherlands	10.11
Wijayatunga and Jayalath (2004)	GDP	2001	Sri Lanka	1.21
Hsu et al. (1994)	GDP	1991	Taiwan	0.07
Aiyar (1995)	Proxy	1995	India	0.20
Parikh et al. (1995)	Proxy	1994	India	0.09

4.4 Cost and Quality Relation

4.4.1 Setting the Quality Target

From the previous discussion, it follows that the firm would deliver an optimal quality level if the financial incentive is set on the basis of interruption costs. As all costs, both network and interruption costs are then internal to the firm, the profit maximising incentive would automatically lead to the socially optimal quality level. However, this does not necessarily mean that the firm will be financially sustainable. If all interruption costs were internalised, the firm would likely run a loss, as the initial price would not cover these additionally incurred costs. A practical solution for this problem is to internalise only a part of total interruption costs. This can be done by configuring the incentive scheme such that only the costs corresponding to deviations from some quality target would be incurred by the firm.[11] By setting an appropriate quality target, the firm's financial sustainability can be maintained while the incentive for optimal quality remains. The question then is how to determine the appropriate level for this target.

Setting the quality target can be thought of to consist of two separate problems: How to set the initial target and how to adjust this target over time. Alexander (1996) presents three alternatives to set a quality target. Firstly, it could be based on historical quality levels. The idea is that a firm should attain a similar performance level as it has previously been able to. Secondly, one could observe quality levels in other regions or countries and use these as a benchmark. The argument here is that if other firms can achieve better results, non-performing firms should demonstrate why similar results could not be achieved. Thirdly, one could litigate or negotiate performance goals and establish a gradual movement towards that point during the term of the regulation plan.

The second aspect of the problem is how to adjust the target. Irrespective of the level of the quality target, if the incentive is set on the basis of interruption costs, the firm always has an incentive to provide the optimal quality level. The quality target essentially determines how benefits resulting from better quality decisions by the firm are distributed It makes therefore sense to periodically adjust the target in order to reflect the improvement in quality performance achieved by the firm.[12] Given that the quality target is initially different from the optimum, this implies that the incentive would not be zero even in the case that the firm provides the optimal quality level. Such a fixed target is problematic because of two reasons. Firstly, if the incentive is positive (the firm receives a reward) then consumers will be paying an additional price to the firm. Although such payments would not seem a problem for a short period, in the longer term consumers may question the distributional properties of the quality scheme. The second problem of a fixed target is that the firm may end up paying a penalty even if it provides an optimal level of quality. This may occur if initially the regulator prescribed an realistically high target for the firm.

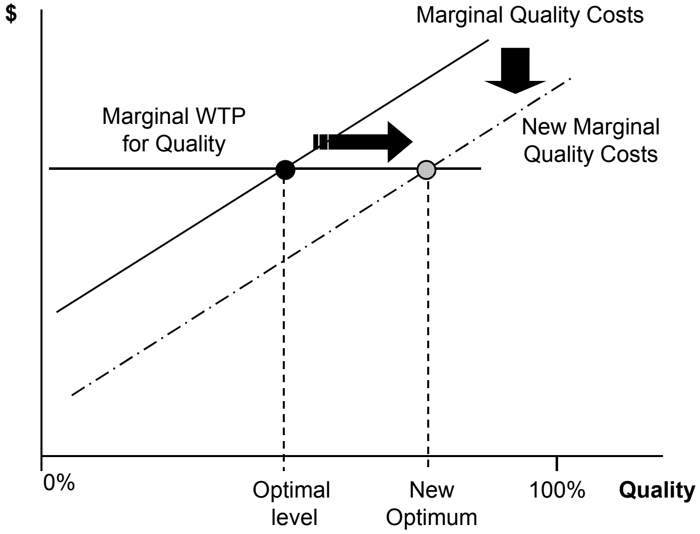


Figure 4-4. Dynamics of optimal quality regulation. WTP = Willingness to Pay.

By tracking actual quality levels over time, at least in theory, the quality target would eventually arrive at the optimal quality level.[13] In reality however, the regulator needs to take into account that the optimal quality level is dynamic i.e. shifts over time. As firms become more productive, they require fewer costs to deliver the same level of quality. Alternatively, they can produce higher quality at the same level of costs. As may be observed from Figure 4-4, productivity improvements results in lower (marginal) costs, and consequently in an increase of the optimum. The intuition here is that as firms become more efficient, they can provide the same quality at lower costs and consequently, the level of quality that they could supply to consumers at given costs should be higher.

The fact that due to quality incentives, cost levels will change, points out to another important observation. The X-factor reflects the regulator's estimation of productivity improvements that the firm could achieve. However, the X-factor should also take into account the fact that the firm's absolute level of costs will change because of choosing a different quality level. Thus, costs would not only change because of changing productivity levels, but also because of differences in quality levels. This interlink should be reflected in the setting of the X-factor. For example, if the firm would need to increase its quality (e.g. if quality is initially below the optimum) then its costs will increase. The regulator should recognise this and allow the firm to (structurally) increase its price in order to finance the additional costs involved in providing a superior quality level. At the same time, it should also be recognised that the firm could increase its productivity and that this will lead to lower costs. Part of the required price increase due to higher quality could thus be financed through higher productivity. But it is not certain whether this will be sufficient to cover the cost increase associated with the quality increase i.e. whether the net effect should be an increase or decrease in prices. As can be observed, there are two separate factors that drive costs namely the productivity level and the level of quality. Ideally, both potential improvements in productivity and quality would need to be captured by the X-factor.

4.4.2 Spatial Variations

The main condition for optimal quality is that costs and consumer demand for quality are equal at the margin. In practical terms, as was shown earlier, this means that the quality incentive is set at a level such that the profit maximisation incentive of the firm reflects the marginal changes in both costs to provide the quality and the associated interruption costs. However, even if from the system point of view quality is optimal, this does not necessarily mean that quality for each individual consumer is also at an optimal level. The reason for this is that both network costs and interruption costs are not uniformly distributed across the network but will vary as a function of the location in the network. Due to these spatial variations, the optimal quality for an individual consumer may not necessarily coincide with the average optimal quality level. This problem can be demonstrated by means of the following model.

Assume that the firm has information about the costs to supply quality to an individual consumer i and that these costs are given by:

$$c^i = z^i \cdot f(q) \tag{4-21}$$

Here, c stands for costs, q for quality, and z is a variable that varies with the specific circumstances of a given consumer. A higher value for z implies that it requires higher costs to provide quality and conversely. For example, for a rural consumer, the variable z would have a higher value than for an urban consumer, which reflects the higher costs involved in serving the former category of consumer. Suppose now that the regulator imposes an optimal quality incentive scheme and does so at the level of the individual consumer. The incentive is linear and defined as:

$$\Phi(q^i) = (q_0 - q^i) \cdot \varphi^i \tag{4-22}$$

Here, q_0 is the quality target and φ^i is the marginal incentive and reflects the interruption costs incurred by consumer i. With quality internalised, the firm incurs an effective level of costs per consumer of:

$$z^i \cdot f(q) + (q_0 - q^i) \cdot \varphi^i \tag{4-23}$$

Assuming that the revenue collected for a given consumer remains constant, profit maximisation implies minimisation of costs for that consumer:

$$z^i \cdot \frac{df^i}{dq^i} = \varphi^i \tag{4-24}$$

The firm would choose a quality level of service for each individual consumer based on the above condition. Then, profits would be maximised. As can be observed, given that φ^i is fixed (by the regulator), there will be a trade-off between quality and complexity: At higher complexity levels (larger values for z), the firm will choose to operate at a lower quality level. That is, consumers where the costs to provide quality are higher, will be provided with a relatively lower level of quality. Conversely, for consumers where the costs to provide quality are lower, the firm will provide a higher quality as this will increase profits.

Complexity is likely to change as a function of the location in the network. For example, it typically requires more costs to produce high quality in rural and sparsely populated areas than in more densely populated and urban areas. In rural areas, distances between consumers are larger and this leads to higher costs per consumer. Furthermore, because the average length of network lines per consumer is larger, there is an increased probability of interruptions. Also, it may require more time to restore supply due to longer travelling distances for repair crews. Consequently, the costs of providing quality to rural consumers are higher and this will be reflected in a lower optimal quality level.

Due to differences in demand and supply conditions (reflected by φ and z), the optimal quality level will vary as a function of the location in the network. In principle, if the incentive scheme were to reflect these differences, the firm would offer a different (but optimal) quality level to each consumer, based on both its own quality preferences and the costs to provide that quality. However, quality differentiation – even though desirable from an efficiency point of view – may not be considered equitable. Some consumers (e.g. in remote areas) may be supplied with very low quality which, even though this may be economically efficient, may not be considered sufficient in particular when taking into account that other consumers (e.g. in urban areas) enjoy very high quality.

To deal with the equity problem, the regulator could impose a minimum standard on top of the incentive scheme. As previously discussed, however, the firm does not necessarily have to meet a minimum standard. It could choose to violate the standard if the cost savings are higher than the associated penalty. Therefore, the minimum standard would need to be binding i.e. the penalty level should be sufficiently high to make sure the firm always complies with the minimum. Such binding standards would make sure that all consumers are guaranteed at least a certain level of quality. However, even if this were possible, a binding standard would have the problem of distorting the incentives provided by the optimal scheme. This can be demonstrated using the following model.

Assume that the minimum standard is binding i.e. the firm always delivers the minimum quality level (q^M), irrespectively. If the optimal quality level (q_*) is higher than or equal to the imposed minimum standard (q^M), then the firm will always provide the optimum level as this leads to profit maximisation. This group of consumers is denoted as high-quality consumers (with index *H*). In the case that the optimal quality is lower than the minimum, then the firm will need to increase quality in order to comply with the minimum standard. The consumers that fall under this group are denoted as low-quality consumers (with index *L*). For the low-quality consumers, the firm needs to increase quality but this will come at higher costs. Thus, a minimum standard will increase the firm's costs to a level higher than incurred at the optimum. Clearly, the higher the minimum standard, the higher these additional costs will be. This implies that applying a minimum standard would require the regulator to allow an increase in price to finance these additional costs. Otherwise, the firm's financial health would be negatively affected.

The distorting effects of the minimum standard can partially be mitigated by using geographically differentiated standard. Minimum standards could be reduced for consumers who are located in regions where quality provision is more costly (rural areas). Similarly, higher standards could be applied for consumers located in areas where costs are lower (urban areas). This approach would reflect the general tendency of costs being proportional to the level of consumer density. Areas where consumer density is high would then arguably be provided with a higher level of quality. This phenomenon can also be noticed in practice. Figure 4-5 shows the relation between consumer density and the average level of quality for a number of distributors worldwide. As can be observed, there is a clear trend of increasing quality with higher consumer density levels. Roughly speaking, distributors serving urban-type areas are located on the right side of the figure. Distributors that serve rural-type areas are located on the left side. Distributors serving mixed areas are located in the middle of the figure.

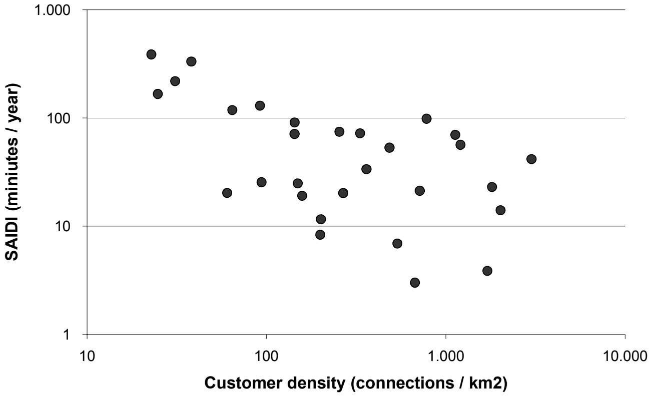


Figure 4-5. Relation between consumer density and quality level (logarithmic scale). See Annex III for a full description of the data.

An assumption made so far is that the incentive level is set for each consumer separately. In reality, the regulator will not likely have information about the quality preferences for each individual consumer. Furthermore, imposing an incentive scheme at the level of this individual consumer would lead to very high administrative costs. Usually, the incentive scheme and the corresponding value for the financial incentive are therefore applied on the basis of the average consumer. This has the effect that the quality offered to an individual consumer may not necessarily reflect his or hers own quality preferences. If the consumer's willingness to pay for quality is higher than implied by the incentive, the quality level will be lower than the theoretical optimum for that consumer and vice versa. To deal with this problem, the regulator could vary the incentive level per consumer class. For consumers that demand higher quality, the incentive level can then accordingly be set higher. This mitigates the distorting effects but does not completely remove them, as there will also be differences in interruption costs amongst consumers belonging to the same class.

4.4.3 Quality Feedback Time

The quality controls that were discussed in this chapter take a long-term perspective. The assumption is that the firm anticipates the impact of future changes in quality demand on its costs and consequently the outcome of the quality control. However, the electricity network industry is notoriously capital intensive and consequently has long asset lives. This long-term character is also reflected in a feedback time between cost decisions and quality outcome. It may take some time before the effects of today's cost decisions are reflected in a change in quality. If this feedback time is long and the firm discounts future incentives at a high rate, then the firm's managers may engage into actions to increase short-term profits even if this would lead to adverse quality effects on the longer term. This may be particularly true if the period for which management is appointed is short relative to the quality feedback time i.e. when the period between the cost decision and the outcome in terms of the incentive is too long.

Generally speaking, the root of the problem lies in the time lag between a change in the firm's inputs (costs) and its effect (adjusted quality). An implicit assumption by the regulator is that the firm makes cost decisions that result in a desirable (e.g. minimum or optimal) quality level term. In principle, the firm is free to choose its own spending levels and consequently make a quality decision. If the time lag between costs and quality is long, however, there is a risk of long-term quality degradation. The intuitive solution for this problem would be to assure that the firm maintains a certain minimum level of spending. This minimum target could be determined by observing past, proposed, and actual expenditure levels and critically examining differences between them. It is also possible to monitor a more detailed level of expenditure, for example by equipment type or function. An alternative approach may be to closely monitor the firm's quality performance e.g. by monitoring the performance of individual network components and prescribing corrective actions in case that a quality decrease is noticed. Periodically, the firm would have to present a detailed report to the regulator containing an overview of the performance for each network component as well as actions planned to guarantee future performance. This enables the regulator to perform a detailed analysis of the health of the network and pick up performance trends well in advance. Finally, the regulator could evaluate (or even prescribe) the firm's business processes and perform audits to verify whether the firm is acting in line with them. In this way, the regulator wants to assure that the firm has established all necessary processes required to guarantee long-term reliability, and is acting in accordance with the processes. This assessment can be done through surveys or by comparing business processes between firms.[14]

4.5 Conclusions

4.5.1 Synthesis

Regulation of network quality has been the central theme of this chapter. The issue of optimal quality and the problems involved in reaching this optimum have been analysed. As can be observed, establishment of an effective quality regulation system is hindered by lack of information about consumer demand for quality and about the relation between cost and quality. The above two information problems have important consequences for the establishment of integrated price-quality approaches.

In order to provide an optimal incentive, the regulator needs to internalise consumer interruption costs into the firm's decision-making process. In acquiring information about interruption costs, the regulator does not only face an informational asymmetry between itself and the firm, but also one between itself and the consumers. Obtaining information about quality demand, approximated by interruption costs, can be a difficult undertaking due to the many factors that can influence these costs.

The second problem is concerned with the cost-quality relation. Costs and quality can vary both in the spatial and in the temporal dimension. The spatial problem leads the optimal quality level to vary as a function of the location in the network. If the firm were to optimise accordingly, then potentially large variations in quality levels may arise. To deal with this presumably inequitable outcome, minimum standards can be imposed but these distort the optimal quality incentives. On the other hand, the occurrence of a time lag between cost decisions and quality creates uncertainty whether quality levels will be optimal in the long-term. To counteract potential short-term optimisation by the firm, minimum spending levels could be prescribed either directly or indirectly. However, this would not be in line with the general idea of regulating on the basis of incentives rather than direct regulatory intervention.

In the design of an integrated price-quality system, the regulator will need to take into account the two informational problems. Such an integrated system would feature an X-factor that incorporates the firm's improvement potential in both the productivity and the quality sense. Under the totex approach, this could be achieved by incorporating quality into the benchmarking analysis. Rather than only observing the firm's actual costs, the regulator would then also take into account the effects of the firm's cost decisions on quality and interruption costs experienced by consumers. However, such a benchmarking analysis is not likely to be conducted at the level of the individual consumer. Rather, it would be performed at the system level and thus ignore possible spatial differentiation in costs and quality demand. Also, with respect to the temporal aspect of the cost-quality relation, it is questionable whether the benchmark would be effective in detecting whether cost decisions will generate a sustained optimal quality level in the longer term.

Under the building blocks approach, spatial and temporal problems would in principle be dealt with as here, the regulator effectively prescribes the required spending level of the firm (at least with respect to investments). If, for each investment, the regulator could assess the quality level provided to individual consumers and the associated costs, he could set a target for the firm's investment levels such that consumers are guaranteed some minimum quality level whilst on the overall network level, quality levels are as close as possible to the social optimum. Furthermore, this excludes the risk of unexpected quality degradation in the longer term. An important assumption here is that the regulator is able to set the appropriate investment target. This requires the ability to predict how the firm's choice regarding its investments would affect the level of quality provided to its consumers. Solving this problem is crucial for designing an effective price-quality system based on the building blocks approach.

4.5.2 The Way Forward

The next two chapters are dedicated to the two integrated approaches for price-quality regulation. Chapter five develops a model for integrated cost-quality benchmarking as part of the totex approach. Chapter six deals with an integrated approach for cost-quality assessment of the firm's investment proposals under the building blocks approach.

Notes

[1] This chapter is based on Ajodhia and Hakvoort (2005).

[2] In the extreme case, some unsatisfied consumers may turn to self-provision. Although the cost of self provision is likely to be costlier than to procure from the monopoly firm, if the quality level is very low, some consumers may find it worthwhile to self-provide anyhow. See also section 4.3 where the use of the costs of private generation as a proxy for the willingness to pay for network reliability is discussed.

[3] See Ajodhia et al. (2006) for a discussion of the presence of such spill-over effects in Italy.

[4] This problem may be particularly relevant in case of major quality problems. Consider the example of recent major British railway incidents which were the result of systematic under-investments in the rail network. However, the public debate was primarily focused on the question whether the policy of privatising the British Rail Industry had been correct or not (BBC 2002).

[5] For some so-called Force Majeure events, breaching the standard does not lead to a financial penalty. Examples of Force Majeure events are faults resulting from exceptional weather, natural disasters, sabotage, war, or terrorist attacks. Majeure events are excluded from the scope of the minimum standard (as well as from other types of quality controls). See for example (CEER 2001) for a listing of possible events that could be classified as Force Majeure.

[6] See Annex I for an overview of reliability indicators in electricity distribution.

[7] For simplicity, the effect of quality changes on demand is ignored. In this case, quality changes would lead both to a direct change in demand (more interruptions would result in less consumption) and an indirect effect (depending on the quality performance, the price would be adjusted and this would influence demand). If quality changes are small, these effects are small and are therefore further ignored.

[8] The idea of setting the incentive under the quality scheme on the basis of interruption costs is, among others, also proposed by Rivier et al. (1999), Williamson (2001), DTE (2002b), and Rivier and Gómez (2003).

[9] Full internalisation of interruption costs will cause financial sustainability problems. This can be solved by setting an appropriate quality target. See also section 4.4.1.

[10] Amounts in local currency have first been inflated to 2004 levels, and then converted to US Dollars using the average exchange rate for 2004. Exchange rates were obtained from the CIA World Factbook, inflation data were obtained from the IMF. These are available at respectively www.cia.gov and www.imf.org. In case the year of the study was not available, the year of the publication has been assumed to be the year of the study.

[11] Conceptually, setting a quality target would be the same as choosing an appropriate value of q_0 from Equation (4-1).

[12] An improvement does not necessarily imply an increase in quality. If initially quality levels were too high i.e. higher than the optimum, a reduction in quality (as well as the corresponding decrease in costs) would be classified as an improvement.

[13] This strategy would be similar to the incremental surplus scheme proposed by Sappington and Sibley (1988).

[14] The risk management survey performed by Ofgem (2002) is an example of this approach.

5

Integrated Price-quality Benchmarking

5.1 Introduction

5.1.1 Background

The previous chapters developed two approaches for integrated price-quality regulation under the totex and building blocks frameworks. This chapter explores the former approach i.e. the use of integrated cost-quality benchmarking analysis for setting the X-factor under totex. This chapter's objective is to develop an integrated benchmarking model and to evaluate this model in the light of the informational problems related to quality regulation: Quality measurement, quantifying consumer demand for quality, and uncertainty about the cost-quality relationship.

5.1.2 Chapter Outline

The benchmarking models used in this chapter are based on Data Envelopment Analysis (DEA). Section two starts by presenting an overview of this methodology and develops two models for incorporating quality into DEA. Section three applies these two models to a data sample of UK and Dutch firms. Section four develops and evaluates a methodology for translating efficiency scores from integrated benchmarking into the X-factor.

5.2 Integrated Benchmarking

5.2.1 Data Envelopment Analysis (DEA)

DEA is a non-parametric mathematical programming approach to productivity frontier estimation. The general idea of DEA is to measure a firm's productivity performance by observing its distance to the productivity frontier which is constructed on the basis of the best performing firms (peers) in the given data sample. According to Coelli et al. (1998), the foundation for DEA was laid by Farrell (1957), who in turn drew upon the work of Debreu (1951) and Koopmans (1951).[1] Farrell (1957) defined a simple measure of productive efficiency that could account for multiple inputs, and easily be generalised for multiple outputs. He claimed that the productivity of a firm consists of two components. Firstly, technical efficiency, which reflects the ability of a firm to obtain maximal output from a given set of inputs, and secondly, allocative efficiency, which reflects the ability of a firm to use the different inputs in optimal proportions, given their respective prices.[2] These two measures are then combined to provide a measure of total productive efficiency.

Farrell (1957) illustrated his ideas by using a simple example involving firms that utilise two inputs (x_1 and x_2) to produce a single output (y), under the assumption of constant returns to scale i.e. output change is proportional to input change across all levels of inputs. The unit isoquant of the fully efficient firm is represented by SS' in Figure 5-1. This efficient isoquant represents the minimum amounts of inputs x_1 and x_2 needed to produce one unit of output y. Under the assumption that the efficient isoquant is known, the technical efficiency of a given firm can be calculated. A given inefficient firm (i.e. one that is not located on the curve SS') uses a certain quantity of inputs, defined by point P, to produce a unit of output. The technical inefficiency for that firm is represented by the distance QP, i.e. the amount by which all inputs could be proportionally reduced without a reduction in output. This is usually expressed by the

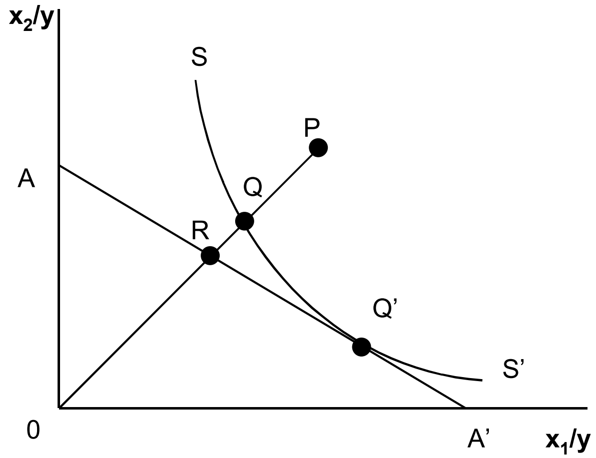


Figure 5-1. Technical and allocative inefficiency. Source: Coelli et al. (1998).

ratio QP/OP, which represents the percentage by which all inputs could be reduced. The technical efficiency level (*TE*) of a firm is most commonly measured by the ratio:

$$TE = OQ / OP = 1 - QP / OP \tag{5-1}$$

This ratio takes a value between zero and one, and provides an indication of the degree of technical (in)efficiency of the firm. A value of one indicates that the firm is fully technically efficient. For example, point Q is technically efficient because it lies on the efficient unit isoquant.

If the relative prices of inputs x_1 and x_2 are known, allocative efficiency i.e. the extent to which an optimal choice between the two different inputs is made, may also be calculated. The input price ratio is represented by the line AA' and the allocative efficiency level (*AE*) of the firm operating at point P is defined by the ratio:

$$AE = OR / OQ \tag{5-2}$$

The distance RQ represents the reduction in production costs that would occur if production were to take place at the allocative (and technically) efficient point Q', instead of the technically efficient, but allocative inefficient, point Q. To sum up, the total economic efficiency level (*EE*) can be defined as:

$$EE = OR / OP \tag{5-3}$$

Here, the distance RP can also be interpreted in terms of a cost reduction. Notice that the product of technical and allocative efficiency provides the overall economic efficiency level, which is also known as 'Farrell efficiency':

$$TE \cdot AE = \frac{OQ}{OP} \cdot \frac{OR}{OQ} = \frac{OR}{OP} \equiv EE \tag{5-4}$$

Notice further that all measures are bound by zero and one, and that overall (Farrell) efficiency has the property of multiplicative separability into input-allocative and technical efficiencies.

5.2.2 DEA as a Linear Program

The above efficiency measure assumes availability of information about the efficient productivity frontier i.e. the unit isoquant of the fully efficient or 'peer' firm. However, this is rarely the case, so that the best-practice unit isoquant must be estimated from sample data. Farrell suggested the use of either a non-parametric piecewise-linear convex isoquant constructed in such a way that no observed point should lie to the left or below it (refer to Figure 5-2), or a parametric function being fitted to the data, again in such a way that no observed point should lie to the left or below it. This concept of the piecewise linear convex hull (the data envelope) approach to frontier estimation – which forms the basis of DEA as it is known today – was considered by only a handful of authors in the two decades after Farrell published his ideas. Authors such as Boles (1966) and Afriat (1972) suggested mathematical programming methods which could achieve the task, but their method did not receive wide attention until a paper by Charnes, Cooper, and Rhodes(1978) inserted Afriat's methodology within a standard operations research framework, and coined the term 'Data Envelopment Analysis'. Since then, there have been a large number of papers extending and applying the DEA methodology.

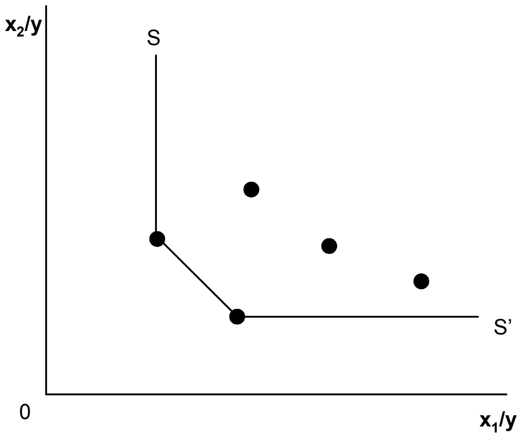


Figure 5-2. The Data Envelope (Coelli et al. 1998).

The basic idea of DEA, as proposed by Charnes et al. (1978) is to solve the efficiency score for each firm based on a linear program formulation. This is done as follows. Consider a data sample consisting of N firms with each K input and M output factors. The vector $\mathbf{x}_j$ represents the inputs used by firm j to produce a set of outputs $\mathbf{y}_j$. Suppose now that $\mathbf{u}$ is an $M \times 1$ vector of output weights and $\mathbf{v}$ a $K \times 1$ vector of input weights. In that case, the general measure of efficiency is provided by:

$$\frac{\mathbf{u}^{\mathrm{T}}\mathbf{y}_j}{\mathbf{v}^{\mathrm{T}}\mathbf{x}_j} \tag{5-5}$$

That is, efficiency is defined as the weighted ratio of outputs over inputs. By definition, efficiency is a scalar between zero and one, which denotes no and full efficiency respectively. The efficiency for firm j can now be calculated by finding appropriate values for $\mathbf{u}$ and $\mathbf{v}$. This requires maximisation of all efficiency ratios under the constraint that these are equal or less than one. This can be formulated as the following optimisation problem:

$$\begin{aligned} &\max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^{\mathrm{T}}\mathbf{y}_j}{\mathbf{v}^{\mathrm{T}}\mathbf{x}_j} \\ &\textit{subject to} \\ &\frac{\mathbf{u}^{\mathrm{T}}\mathbf{y}_k}{\mathbf{v}^{\mathrm{T}}\mathbf{x}_k} \le 1, k = 1 \ldots N \\ &\mathbf{u}, \mathbf{v} \ge 0 \end{aligned} \tag{5-6}$$

Solving this problem, however, yields an infinite number of solutions. This can be overcome by adding an additional constraint:

$$
\begin{aligned}
&\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^{\mathrm{T}}\mathbf{y}_j \\
&\textit{subject to} \\
&\mathbf{v}^{\mathrm{T}}\mathbf{x}_j = 1 \\
&\mathbf{u}^{\mathrm{T}}\mathbf{y}_k - \mathbf{v}^{\mathrm{T}}\mathbf{x}_k \le 0, k = 1 \ldots N \\
&\mathbf{u}, \mathbf{v} \ge 0
\end{aligned}
\qquad (5\text{-}7)
$$

Using duality in linear programming, this can then be written down in the common form for the DEA problem:

$$
\begin{aligned}
&\min_{\theta,\lambda} \theta \\
&\textit{subject to} \\
&-\mathbf{y}_j + \mathbf{Y} \cdot \lambda \ge 0 \\
&\theta \cdot \mathbf{x}_j - \mathbf{X} \cdot \lambda \ge 0 \\
&\lambda \ge 0
\end{aligned}
\qquad (5\text{-}8)
$$

The matrices **X** and **Y** represent respectively the input and output data space that consist of the individual input and output vectors $\mathbf{x}_j$ and $\mathbf{y}_j$ for all N firms. The optimisation problem needs to be run for each firm and results in its efficiency score θ.

An intuitive interpretation of the DEA formulation is that of measuring the distance to a multi-dimensional productivity frontier. This frontier is constructed by enveloping all efficient input and output combinations. The efficiency measure is then obtained by measuring the distance between the firm's actual performance against that of its projection (shadow) on the frontier.

The original DEA model, as described above, is based on the assumption of constant returns to scale (CRS). The general idea is that costs change linearly in response to an output change. That is, the efficiency score also reflects inefficiencies arising from choices of operating scale. As the firm's scale may sometimes not be within the control of the firm, it may be necessary to correct the efficiency score for such scale differences. This is done under the variable returns to scale (VRS) model. Here, only firms of similar scale are compared to each other. This is done by constraining the size of λ to unity. The linear problem then has the following modified form:

$$
\begin{aligned}
&\min_{\theta,\lambda} \theta \\
&\textit{subject to} \\
&-\mathbf{y}_j + \mathbf{Y} \cdot \lambda \ge 0 \\
&\theta \cdot \mathbf{x}_j - \mathbf{X} \cdot \lambda \ge 0 \\
&\mathbf{N1}^{\mathrm{T}} \cdot \lambda = 1 \\
&\lambda \ge 0
\end{aligned}
\qquad (5\text{-}9)
$$

Where **N1** is an $N \times 1$ vector of ones. The difference between the CRS and VRS frontiers in the case of a single input and output is shown in Figure 5-3. As can be observed, the VRS frontier provides a more lenient measure of efficiency as perceived inefficiencies resulting from scale differences are now not anymore reflected in the efficiency score. This is done by comparing firms only to peers, which operate at a similar scale as themselves. The difference between the VRS and the CRS efficiency score then acts as a measure for scale inefficiency. In this way, the overall inefficiency can be decomposed into scale inefficiency and pure technical inefficiency.

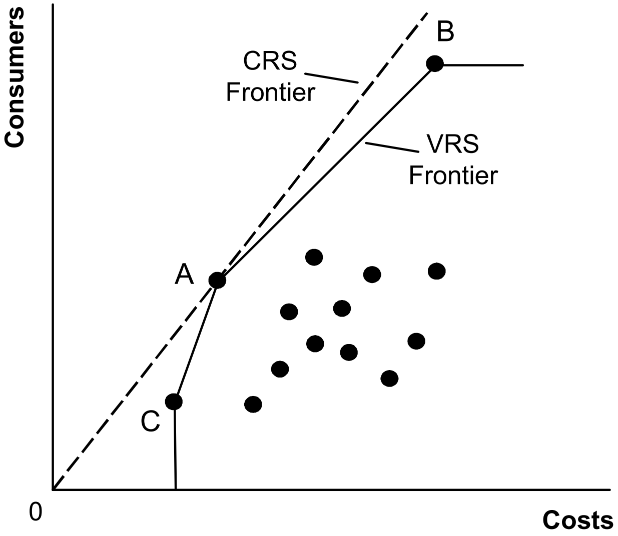


Figure 5-3. Difference between the CRS and VRS frontier under DEA.

The regulatory preference for a CRS or VRS model depends on mainly two issues. Firstly, in the case that firms have no control over their scale (e.g. due to natural or legal constraints), the choice for VRS may be more appropriate. In that case, firms would not be punished for inefficiencies resulting from improper scale selection over which they have no control anyhow. Conversely, in the case that firms can influence their scale (e.g. through mergers), then the regulator may provide an incentive for firms to choose an optimal scale by using CRS.

The second issue related to the CRS or VRS choice is more pragmatic. When moving from CRS to VRS, data requirements of the DEA model increase. Simply stated, a VRS analysis can be thought to consist of a number of CRS analyses, each carried out for a subset of similarly sized firms. This implies that in order for the results to maintain sufficient discriminative power, the sample size should be large enough. Otherwise, the probability of a certain firm being unjustly classified as fully efficient due to a lack of comparators increases. This follows from the property of DEA where firms for whom no suitable comparators can be found, are assumed fully efficient by default. Consider for example Figure 5-3 which shows the CRS and VRS productivity frontier for a given sample of firms and taking into consideration only a single input (costs) and a single output (number of consumers).

The CRS frontier consists of firm A alone as this firm produces the largest amount of outputs for the given inputs. Under the CRS technology, the efficiency of all other firms would be measured against firm A, irrespective of their scale. As may be observed in Figure 5-3 , there is one firm (firm B) that is substantially different from the others as it operates at a much larger scale i.e. serves a larger number of consumers than other firms. Under a VRS specification, this firm B would now be considered as efficient i.e. form part of the VRS frontier as now, scale effects are taken into account. However, this does not necessarily mean that firm B is truly efficient as its efficiency of one may rather be explained by a lack of comparators for firm B i.e. the absence of similarly sized firms within the given data sample.

5.2.3 Incorporating Quality into DEA

There is an increasing interest in the use of DEA by regulators – in particular for setting the X-factor (see also Table 3-1). DEA has the advantage of generating a quantitative measure of efficiency that is intuitive and relatively easy to comprehend: The efficiency score simply reflects the amount by which the firm could improve its performance up to the level of its peers. Furthermore, DEA is capable of handling multiple input and output factors. This takes into account the multi-dimensional nature of the electricity distribution service. DEA also does not require any functional specification of the relationships between these different inputs and outputs. Rather, DEA deterministically constructs the productivity frontier based on observed best-practice performance in multiple input/output dimensions. Furthermore, the fact that DEA is relatively easy to apply gives DEA an advantage over statistical more demanding benchmarking techniques.

Like any other benchmarking technique, the outcome of the DEA benchmark is limited by the choice of model specification. In line with Cubbin (2003), four types of model specification errors can be identified. Firstly, omitting a relevant variable will reduce the accuracy of the analysis. Secondly, adding an irrelevant variable will increase the bias and raise the proportion of falsely attributed efficiency scores of one. Thirdly, since DEA allows both constant and variable returns to scale specification, errors may also occur through using the wrong model while fourthly, specifying input factors as output factors or conversely will influence the outcome of the analysis.

Due to the deterministic character of DEA, it is not possible to test the statistical significance of the obtained outcomes, i.e. verify the validity of the model's assumptions. This also applies to possible errors in the underlying data used in the analysis. The efficiency score is deterministically constructed on the basis of the given sample data. If there are errors in these data, then the efficiency scores will potentially also be wrong. Furthermore, given the interdependency between firms, errors in the data of one firm can potentially drive the efficiency scores of other firms. Regulatory experience suggests that assuring high quality data is an essential part of applying a successful DEA benchmark. This is particularly true in the case that the translation of efficiency score into an X-factor is more direct. Recognition of the limitations of the benchmarking exercise – with respect to both the model specification and the quality of the data – remains imperative. This applies to DEA as well as any other benchmarking method.

Table 5-1. Overview of DEA studies of electricity distribution. *I* stands for input factor and *O* stands for the output factor in the model's specification.[3]

	Model Specification														
DEA Study	**Opex**	**Manpower**	**Totex**	**Investments**	**Capital**	**Transformers**	**Network length**	**Network Loss**	**Sales**	**Peak Demand**	**Customers**	**Service Area**	**Network Size**	**Density Factors**	**Quality**
Hjalmarsson and Veiderpas (1992)		I					I		O	O					
Weyman-Jones (1991)		I				I	I		O		O				
Hougaard (1994)	I	I			I			I	O		O		O		
Pollit (1995)		I				I	I		O	O	O	O			
Førsund and Kittelsen (1998)		I			I			I	O		O			O	
Kumbhakar and Hjalmarsson (1998) - 1		I				I	I		O		O				
Kumbhakar and Hjalmarsson (1998) - 2		I				I	O		I		I				
London Economics (1999)		I					I		O		O				
Scarsi (1999) - 1		I					I				O				
Scarsi (1999) - 2		I					I		O						
Scarsi (1999) - 3	I					I	I		O	O	O				
DTE (2000)			I			O	O		O	O	O				
Pardina and Rossi (2000)		I							I		O	I			
Lo et al. (2001)	I				I	I	I		O		O				
Korhonen and Syrjänen (2002)															O
Pahwa et al. (2002)	I			I		I	I	I	O	O	O				
Resende (2002) - 1		I				I	I		O		O	O			
Resende (2002) - 2		I				I	I				O	O			
Ajodhia et al. (2004)	I						O		O		O				
CEPA (2003) - 1			I				O		O		O				
CEPA (2003) -2							O		O		O				O
Giannakis et al. (2003) - 1	I						O		O		O				
Giannakis et al. (2003) - 2			I				O		O		O				
Giannakis et al. (2003) - 3							O		O		O				I
Giannakis et al. (2003) - 4			I				O		O		O				I

Table 5-1 provides an overview of some DEA studies of electricity distribution. As may be observed, quality is generally not considered in DEA studies. This is an important limitation as excluding quality ignores the fact that costs are not only related to the level of outputs, but also to the quality level at which these outputs are supplied. For a given level of outputs, higher quality will generally lead to higher costs. Thus, in a cost-only DEA model, firms who are providing high quality may potentially be incorrectly classified as less efficient. The following example demonstrates this point.

Let c stand for costs, η is the productivity level, which varies between zero and one (corresponding with no and full productivity, respectively), x is the number of consumers and q is the level of quality that these consumers receive. Assume that there is the following simple linear relationship between costs and quality:[4]

$$c = \frac{1}{\eta} \cdot q \cdot x \qquad (5\text{-}10)$$

Assume that there are two firms A and B who are both fully productive ($\eta = 1$) and serve the same number of consumers ($x_A = x_B$). Assume now that firm A chooses to provide a quality level that is twice as high as firm B ($q_A = 2 \cdot q_B$). Consequently, firm A incurs twice as much costs per consumer than firm B ($c_A = 2 \cdot c_B$). Now, if the efficiency scores for these two firms would be measured as cost per consumer then firm B would have least cost per consumer and would therefore be provided with an efficiency score of one. Firm A would incur twice the costs as firm B and its efficiency score would therefore be set at 50 percent. Thus, even though both firms are equally productive, they would still be assigned different efficiency scores because of differences in their quality policy. Consequently, the X-factor for firm A will be higher than that for firm B. From the (ill-informed) regulator's perspective, firm A has to reduce its costs to the level of firm B as, so the regulator believes, firm A's costs are explained by lower productivity performance. However, in reality firm A cannot improve further as it is already fully productive. The only way to comply with the regulatory target to reduce costs by half, will be to reduce quality to the level supplied by firm B.

If the regulator had also set in place a separate incentive scheme for quality, an additional problem may occur. Suppose that the quality scheme sets a base target equal to the historic quality level and penalises or rewards a quality performance below or exceeding this target, respectively. Firm A, which would have an incentive to cut costs to the level of firm B, would now also be penalised as the cost reduction would lead to a quality decline. Effectively, firm A would be punished for the fact that it historically provided a higher quality level than firm B. Firm B on the other hand, would be able to increase its quality up to the level where the resulting rewards are higher than the required cost increase.

Although the above example is strongly simplified, it highlights the basic problem of, on the one side, not considering quality into the benchmarking analysis and, on the other side, a potential inconsistency that may occur between the price and quality incentive regimes. Such problems can be avoided under an integrated price-quality approach. In that case, the efficiency score (and consequently the X-factor) does not only reflect potential improvements in the cost sense, but also in the quality sense. At the same time, it should be realised that the firm generally uses multiple inputs and produces multiples outputs and quality levels. The benchmarking analysis should take into account the multi-dimensional nature of this input-output relationship. DEA offers the possibility for multi-dimensional modelling. Two main approaches may be identified for incorporating quality. In the first approach, quality is defined in terms of a technical output factor (technical model). The second approach models quality in terms of a cost input (social costs or sotex model). The latter approach, as will be shown next, is generally preferable to the former one when applying integrated price-quality benchmarking. The two models are now developed in more detail.

5.2.4 Technical Model

Under the technical model specification, quality would in principle need to be defined as an output factor. That is, if inputs increase (e.g. more investments are made) then quality output levels would consequently also improve. Thus, modelling quality as an output factor takes into account the fact that higher quality is associated with higher costs. Providing higher quality at given costs would then lead to a higher efficiency score. Generally, however, quality is defined and measured in terms of its inverse, for example by the number or duration of interruptions. In that case, the efficiency score would need to be increased at lower levels of inverse quality (e.g. fewer or shorter interruptions). A technical model therefore needs to specify inverse quality as an input factor: Higher levels of inverse quality would then – ceteris paribus – lead to a lower efficiency score.

There are only a few studies that use quality in terms of the technical model specification (see also Table 2). The most comprehensive study is by Giannakis et al. (2005) who analyse the performance of distribution firms in the UK between the period 1991 and 1999. In their model, three input factors are used namely total expenditures (totex), the total number of interruptions, and the duration of these interruptions. The number of consumers served and the amount of energy delivered are specified as model outputs. The authors find that including quality into the DEA model specification leads to a significant change in efficiency scores as well as in the relative ranking of firms. Generally, they find that firms that score high in a cost-only model are generally not high-quality providers, which suggests that cost-only models fail to capture the quality aspect of the firms' operations. Therefore, they conclude, it is preferable to incorporate quality variables in the cost- models.

Specification of quality in terms of a technical output (or more specifically, inverse quality as a technical input factor) is one step into the direction of integrated price-quality benchmarking. However, there are some problems involved in this approach. Under DEA, a firm is automatically assigned an efficiency score of one in case it scores best on a given input/output combination, irrespective of how it performs in terms of other input/output combinations. Thus, if a certain firm scores very high in the quality sense, it could be considered fully efficient even if it would perform very badly in the cost sense. Conversely, a firm that performs very well in the cost sense but provides a very low quality would potentially also be classified as efficient.

The specialisation problem is demonstrated in Figure 5-4 where an example is given of the productivity frontier constructed on the basis of two input factors: Costs (c) and inverse quality (s) and one output factor (y) which may be the number of consumers served or total energy delivered or any other relevant output. The productivity frontier is presented for two cases. In the first case, the productivity frontier is made up of firms C, D, and E and is represented by the dotted frontier. The second frontier consists of the solid line connecting firms A, D, and B. Here, firms A and B have specialised in costs and quality, respectively, and due to this, now form part of the frontier. Firm B is providing a very high level of quality (that is, a low level of inverse quality *s*) and manages to be located on the frontier even though it incurs very high costs compared to other firms. Conversely, firm A provides very low quality and yet remains on the frontier because its costs are very low. As may be observed, both firms A and B are making a different trade-off between quality and costs but manage to achieve an efficiency score of one under the technical model specification. This demonstrates the specialisation problem where firms can achieve a high efficiency score by providing either a very high or a very low quality level.

The technical model may provide firms with an adverse incentive to specialise in either high quality or low costs. Specialisation increases the efficiency score although this does not necessarily mean that the resulting trade-off between cost and quality is socially desirable. This can be seen as follows. Firms that choose to provide high quality will incur higher private costs but interruption costs will be very low. Firms that opt for low quality on the other hand, will incur relatively low private costs but consumers will experience very high interruption costs. As one moves from a low quality (e.g. firm A in Figure 5-4) towards a high quality (e.g. firm B in Figure 5-4), total social costs will decrease and then increase again once the optimal quality level has been reached. That is, the optimal quality choice is likely to be located somewhere between the two extremes of low and high quality which correspond to cost and quality specialisation, respectively. In principle, it may well be that the existing quality preferences are such that either the very high or very low quality level coincides with the optimal quality. However, this is not generally the case.

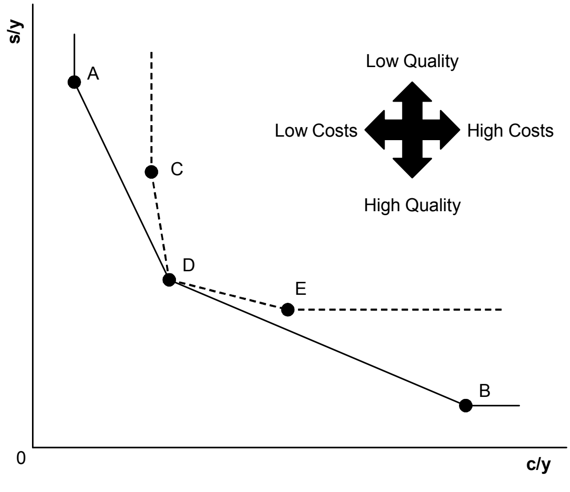


Figure 5-4. Effects of quality and cost specialisation on the positioning of the frontier. Firm A is providing low quality at low costs while firm B is providing high quality at high costs. Both firms are part of the productivity frontier but neither necessarily make an optimal price-quality trade-off.

5.2.5 Sotex Model

Ideally, the efficiency score should reflect the firm's potential for arriving at a more desirable level of total social costs. This can be achieved under DEA by modelling quality in terms of a cost input and, more specifically, by defining social costs (sotex) as the input factor of the DEA model. Sotex includes both the firm's private costs (opex and capex) and the interruption costs incurred by consumers. Having defined sotex as the input factor, the firm's efficiency score will be higher if the firm manages to make a more optimal trade-off between costs and quality i.e. to reduce the level of sotex. On the one side, providing a higher quality level will decrease the level of interruption costs experienced by consumers. On the other side, the firm's own costs will increase. At a given quality level, the sum of these separate elements will be minimised – this will then reflect the optimal quality level. Operating closer to this optimum drives down sotex and potentially increases the efficiency score.

The efficiency score represents the difference between the actual and the desired sotex level – as disclosed by the benchmarking analysis. A score lower than one implies that the firm could potentially decrease its sotex level. This can be done in two fundamental ways. Firstly, the firm could keep productivity fixed and change the level of quality into the direction of the optimum. If initially the firm provides a quality lower than the optimum, it will increase quality. This leads to an increase in network costs but also to an even larger decrease in interruption costs. Similarly, if quality is initially higher than the optimum, a decrease in quality will increase interruption costs but this will be eclipsed by the corresponding decrease in network costs. In both cases, the quality change leads to a net reduction in the level of sotex.

The second way through which the firm can reduce social costs is by becoming more productive. Keeping quality fixed, the firm can increase productivity levels and therefore provide the same level of quality but at lower costs. Alternatively, the firm can produce more units of quality at the same cost. In contrast with quality, productivity maximisation is always a desirable objective as this leads to a reduction in total social costs. A change in quality on the other hand is only desirable if the firm steers into the direction of the optimum quality level.

In short, reducing sotex can be achieved by providing a quality level that is closer to the optimum or by operating at a more productive level (or by a combination of these two). The efficiency score, however, only measures the total improvement potential and does not distinguish between these two separate movements. That is, the efficiency score reflects the potential of the firm to reduce its level of sotex but does not indicate whether this can be achieved by providing a better quality level or by becoming more productive (or through a combination of both). For example, a very productive firm that also provides too high quality may well incur higher sotex than a less productive firm that provides a very low level of quality. The former firm will have a lower efficiency score than the latter but this does not imply that it is providing too low quality. Similarly, the fact that the latter firm has a high sotex score does not imply that it is operating productively.

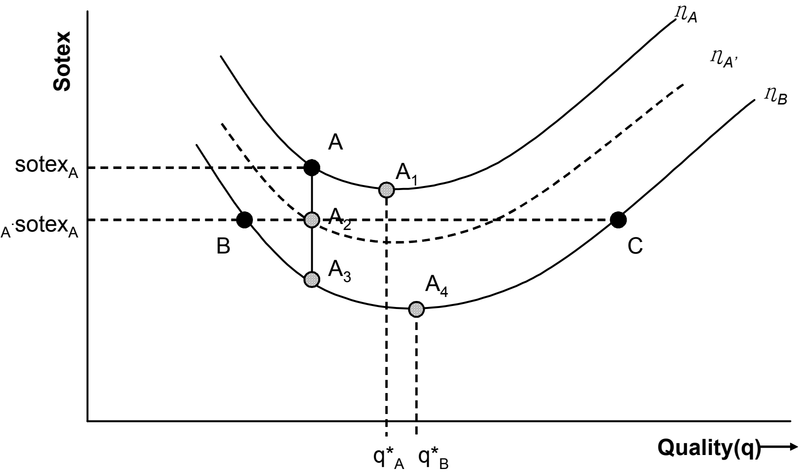


Figure 5-5. Stylised example of the problem of decomposing the sotex efficiency score into a productivity component (represented by the movement from one productivity curve *η* to the other) and a quality component (represented by the movement along a productivity curve).

An important restriction of the sotex model is that it does not disclose information on whether a certain firm is operating below or above the optimal quality level. For a better understanding, consider the example provided in Figure 5-5. Here, three firms A, B and C are considered which are similar except in the level of productivity they operate at and the level of quality that they provide. Initially, firms B and C are more productive than firm A and are located on sotex curve (η_B) below that of firm A. The optimal quality level for firms B and C is therefore also higher than for firm A as at higher productivity levels, the ability of the firm to produce higher quality increases. As the cost to produce an extra unit of quality decreases, the firm could produce a higher quality at the same costs. Thus, higher productivity performance implies an increase in the optimal quality level.

Assume now that a sotex benchmark is performed for the three firms. Firms B and C, which incur less sotex than firm A, will then obtain an efficiency score of one. The efficiency score for firm A will reflect the potential reduction in its sotex levels up to the level of firm B or C which in this case will be equal to A_2/A. Provided with an incentive for reducing sotex, firm A has two main options to improve performance. Firstly, it could keep productivity fixed and operate at a different quality level i.e. one that is closer to the optimum. Firm A could for example move towards point A_1, which is the optimal quality level that corresponds to a productivity level of η_A. As can be seen, its efficiency score will then slightly improve. The second option to decrease sotex is to increase productivity while keeping quality fixed. Firm A could first increase productivity up to the level of $\eta_{A'}$ and operate at point A_2. Then, firm A would incur the same level of sotex as firms B and C and therefore would be considered fully efficient under the benchmark. However, as can be seen, at point A_2, firm A still has potential to improve productivity further towards a level of η_B and operate at point A_3 on curve η_B. In this case, firm A would become the peer for firms B and C, as point A_3 is a better cost-quality pair from the social point of view. In the ideal case, firm A could simultaneously increase productivity towards a level of η_B and then increase its quality performance towards the true social optimum as represented by point A_4. There, it would operate at the highest possible productivity as well as provide a socially optimal level of quality.

At the same time, even though firms B and C initially have an efficiency score of one, further improvement for these firms is also possible. Given that these firms operate at the maximum possible productivity level, the only possibility to reduce sotex levels is to increase quality for firm B and decrease quality for firm C to point A_4 in Figure 5-5. Firm B is initially providing a too low quality level; increasing quality will come at additional costs but this is justified by the decrease in interruption costs. Similarly, firm C is providing a too high quality level and the associated increase in interruption costs due to the quality level decrease is thus offset by the associated reductions in costs.

As can be observed, the sotex model tends to underestimate the true improvement potential both in the productivity and in the quality sense. The underlying source of this problem is the fact that the firm with an efficiency score of one is not necessarily the most productive or the optimal quality-providing firm. In theory, this problem would be absent if the benchmarking sample would include the truly optimal firm i.e. the firm operating on point A_4 in Figure 5-5. In practice however, the data sample is constrained and is not likely to include this truly optimal firm (if it exists at all). This problem is not specific to the sotex model but generally applies to all benchmarking models. Under a cost-only model, the fact that there may be other more productive firms outside the sample also works in favour of the firms evaluated under the benchmarking. In this particular case, however, the absence of a complete benchmarking sample tends to underestimate not only the cost improvement potential, but also the scope for performance improvement in the quality sense.

5.3 Empirical Analysis

5.3.1 Data and Model Specification

To elaborate on the difference between the two DEA models discussed in the previous section, this section applies the models to a sample of 20 firms from the UK and the Netherlands. Ideally, the benchmark model should consider multi-year data in order to reflect the long-term nature of totex. However, due to data constraints, this analysis considers only a single year namely the years

2000 and 2002 for the Netherlands and the UK respectively. The choice for this particular dataset is driven by data availability. For the Netherlands and the UK, audited data is readily available in the public domain. This assures a high level of data quality. A summary of the data is presented in Table 5-2, the full dataset is contained in Annex VII.

Table 5-2. Summary of data used for the benchmarking analysis.

	Unit	Average UK	Average The Netherlands
Opex	USD(000)	131,652	36,075
Depreciation	USD(000)	49,427	15,262
SAIDI	min/year	73	28
CML	min/year	132,597,499	10,964,566
Interruption Costs	USD(000)	15,104	1,327
Consumers	#	1,843,624	385,934
Energy	kWh	20,960,000	4,848,277

Three DEA models have been applied to the dataset; these are shown in Table 5-3. The first model, labelled TOTEX, is a traditional DEA model for analysing cost-only efficiency. The second model (CML) corresponds to the technical model specification and includes (inverse) quality by modelling total customer minutes lost (CML) as a technical input factor. The third model (SOTEX) is based on the idea of using total social costs (sotex) as an input that is defined as the sum of totex and consumer interruption costs. Previous DEA studies that considered quality have only included quality in terms of a technical output. As far as known, a sotex model has not yet been previously applied.

Table 5-3. DEA model specification: Definition of input and output factors.

	TOTEX model	CML model	SOTEX model
Totex	Input	Input	
Customer Minutes Lost		Input	
Sotex			Input
Energy Delivered	Output	Output	Output
Nr. of Consumers	Output	Output	Output

For each model, the energy delivered and the number of consumers served are defined as output factors – this is consistent with the DEA specification generally used for electricity distribution (see also Table 5-1). This output specification is also supported by Allas and Leslie (2001) who find that 85 percent of costs in electricity distribution tend to vary with the number of consumers served and the units of energy delivered.

Furthermore, all DEA models use input-orientation i.e. outputs are considered fixed and efficiency scores depend on the extent by which input factors can be reduced to arrive at frontier performance. Given the relatively small sample size, constant returns to scale (CRS) is assumed. This implies that the efficiency score may partially be driven by scale inefficiencies.

Totex is defined as the sum of reported opex and depreciation costs. Due to data constraints, the cost of capital element has not been included in the totex definition. CML is defined as the cumulated minutes of interruption for all consumers served by the distribution firm and has been calculated by multiplying the SAIDI indicator by the total number of connected consumers. A lower level of CML thus indicates a higher level of quality. For calculating interruption costs, a modified version of the GNP proxy method has been used. Under this method, which is for example used by Telson (1975), interruption costs per kWh are estimated as the annual GNP divided by the total units of electricity consumed in the corresponding year. The basic idea is that during an interruption, there is a loss to society that can be approximated by the production that would have been realised if the interruption had not occurred. Total interruption costs can then be derived by multiplying the costs per kWh by the total amount of energy not supplied in the given year. In this particular case, interruption data was limited to information about SAIDI performance. Consequently, the GNP proxy method has been adapted to take into account this data constraint. Here, the assumption is that in a given year, the total time during which no contributions to the national production (GNP) have been made is equal to SAIDI hours per consumer. Average interruption costs per hour and per consumer are thus given by:

$$\frac{GDP}{(8760 - SAIDI) \cdot \sum N_j} \tag{5-11}$$

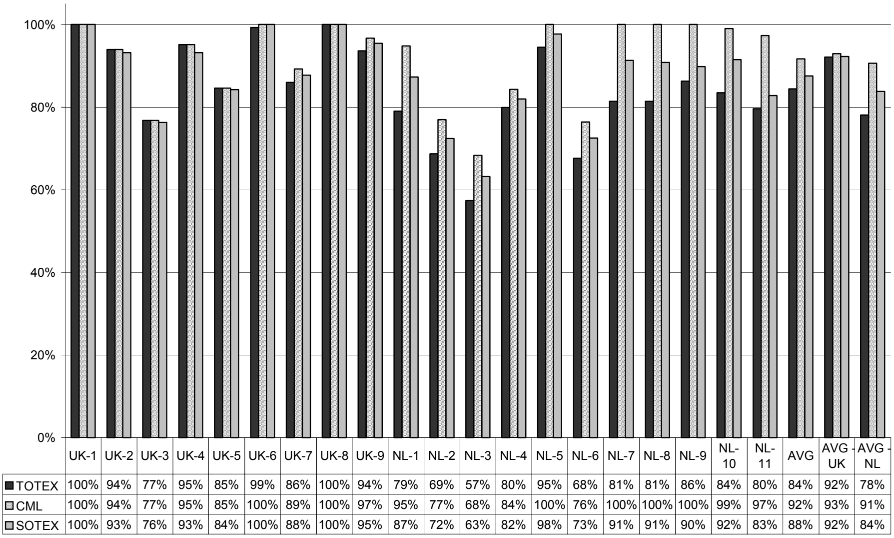
Where 8760 stands for the total number of hours in a year (leap years are ignored) and N_j stands for the total number of consumers served by each distribution firm *j*. Interruption costs (*ic*) for each individual firm can then be given by:

$$ic_j = \frac{GDP}{(8760 - SAIDI) \cdot \sum N_j} \cdot SAIDI_j \cdot N_j \tag{5-12}$$

The advantage of the GNP method is that its data requirements are rather limited and that the data needed is readily available for both the UK and the Netherlands. Furthermore, the derived figures are comparable across countries as the same methodology has been used. Some notes regarding the limitations of the GNP proxy method are in place, however. The interruption cost estimation is a highly aggregated one and does not distinguish between variations in the GDP contribution per geographical region or consumer type. Furthermore, a uniform distribution of production over time is assumed. In reality, production will be time dependent e.g., businesses typically do not operate during certain hours of the day and during certain days in the week. In summary, although the GNP proxy method is not the most accurate, it is preferred in this case for its relative simplicity. In section 5.4.2, the impact of inaccuracies in the interruption cost estimations will be assessed.

5.3.2 Modelling Results

Figure 5-6 provides an overview of the results for each DEA model. The efficiency scores are bound between zero and one and are presented for each (anonymous) firm.



	UK-1	UK-2	UK-3	UK-4	UK-5	UK-6	UK-7	UK-8	UK-9	NL-1	NL-2	NL-3	NL-4	NL-5	NL-6	NL-7	NL-8	NL-9	NL-10	NL-11	AVG	AVG - UK	AVG - NL
■ TOTEX	100%	94%	77%	95%	85%	99%	86%	100%	94%	79%	69%	57%	80%	95%	68%	81%	81%	86%	84%	80%	84%	92%	78%
▨ CML	100%	94%	77%	95%	85%	100%	89%	100%	97%	95%	77%	68%	84%	100%	76%	100%	100%	100%	99%	97%	92%	93%	91%
□ SOTEX	100%	93%	76%	93%	84%	100%	88%	100%	95%	87%	72%	63%	82%	98%	73%	91%	91%	90%	92%	83%	88%	92%	84%

Figure 5-6. Results of DEA benchmark. UK=United Kingdom, NL= the Netherlands, AVG=Average.

Model TOTEX is the base model and shows the efficiency in the area of total expenditure only. As can be observed, efficiency scores are generally higher for UK firms. This implies that Dutch firms are operating at higher cost levels (per unit of output) compared to their UK counterparts. This does not necessarily imply that Dutch firms are less productive as (part of) their higher costs may be explained by higher quality. Indeed, SAIDI levels in the Netherlands are substantially lower than in the UK: Whereas the average SAIDI for UK firms is 73 minutes, average SAIDI for Dutch firms is only 28 minutes (see also Table 5-2). Higher quality generally drives up costs and, as quality is not considered in the model, this may (partially) explain the lower efficiency scores for Dutch firms.

To investigate this issue, the efficiency scores have been regressed on the level of quality (SAIDI). The regression (see Figure 5-7) shows only a weak correlation between the efficiency score and provided quality – as indicated by the low correlation factor of 0.1753. However, this result should be interpreted with care as the DEA score is bound between zero and one. Under DEA, a given firm's efficiency score cannot increase beyond one, irrespective how well it performs on a particular input/output combination. At the same time, a better input/output performance does not necessarily imply a proportional decrease in the efficiency scores of the other firms. Furthermore, the models only include energy delivered and consumers served as output factors whilst there may be other cost drivers that are not captured here.

Notwithstanding the low correlation factor, efficiency scores for high-quality firms increases substantially under the CML model: The average efficiency score for Dutch firms increases from 78 percent to 91 percent. For UK firms, the average score remains stable around 93 percent. The increase in average scores is a normal consequence of including additional factors into a DEA model as this reduces the explanatory power of the model. To overcome this problem, Giannakis et al. (2005) suggest considering the relative ranking of firms rather than their absolute efficiency score. As shown in Table 5-4, the ranking for Dutch firms generally improves when moving to the CML model. More particularly, three Dutch firms are now also located on the frontier i.e. have a ranking of one. Thus, inclusion of quality as a technical output seems to favour Dutch firms to a large extent.

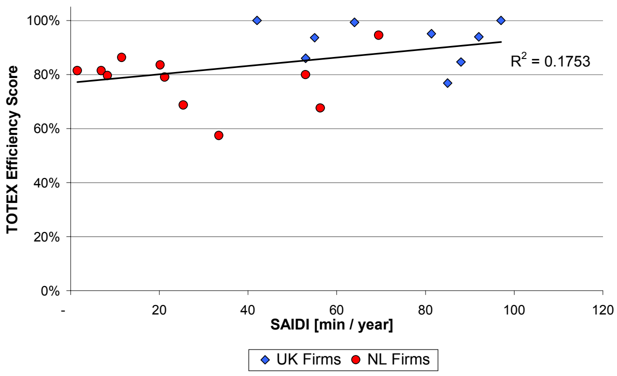


Figure 5-7. Correlation between quality performance and efficiency score under the TOTEX model.

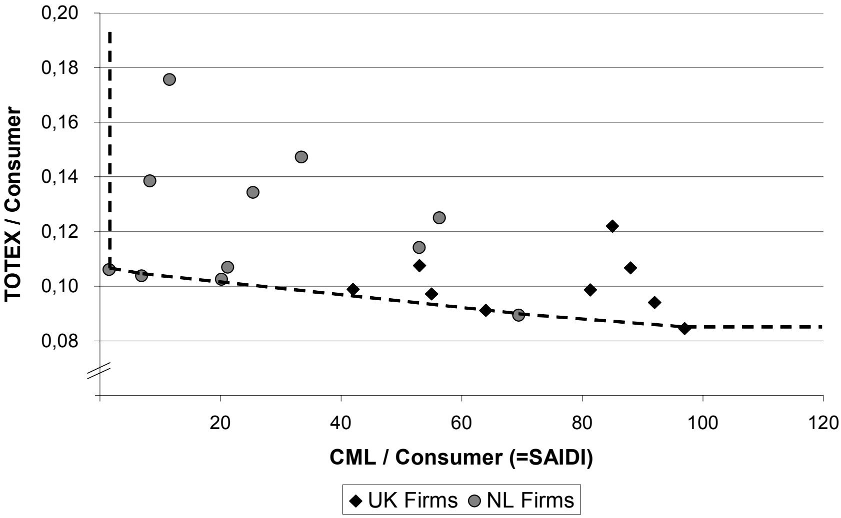


Figure 5-8. Graphical representation of the DEA frontier under the CML model.

As was discussed earlier, efficiency scores under the CML model may be driven by specialisation. To investigate this issue, it is helpful to consider Figure 5-8 where the frontier with respect to the totex per consumer and CML per consumer ratios is shown. Roughly speaking, the sample can be divided into two subsets namely one of firms operating at the high quality-costs equilibrium and another of firms that operate at the low quality-cost equilibrium. These two subsets correspond to Dutch and UK firms, respectively. Dutch firms score particularly well on the CML/consumer ratio but generally have higher per-unit costs than their UK counterparts. Conversely, UK firms manage to achieve a high efficiency score due to their superior cost performance.

Cost and quality specialisation by UK and Dutch firms, respectively, leads both to be classified as efficient, but this does not necessarily imply that an optimal trade-off is made between cost and quality. Thus, a high score under the CML model does not necessarily mean that Dutch firms are productive i.e. produce this high quality at lowest possible costs. Similarly, the fact that UK firms have high efficiency scores does not mean that they are providing an optimal level of quality at the given low level of costs. This specialisation problem is absent under the SOTEX model where total social costs act as the only input factor. Then, firms that specialise in either costs or quality would not automatically be assigned a high efficiency score. Rather, the efficiency score will reflect to what extent the firm manages to minimise the sum of these two, i.e. make an optimal trade-off between costs and quality.

As can be observed from Figure 5-6, when moving from the CML to the SOTEX model, the average efficiency score for Dutch firms decreases from 91 percent to 84 percent – note that this is still higher than the original score of 78 percent under the TOTEX model. Similarly, the ranking under the SOTEX model is more in line with that under the TOTEX one. This observation can be explained as follows. As Dutch firms provide higher quality than their UK counterparts do, they also incur less additional interruption costs under the SOTEX model compared to the TOTEX model. Hence, Dutch firms experience a lower increase in the input factor (outputs are the same under both models) and therefore have a relative improvement in efficiency scores. However, UK firms still determine the frontier implying that these firms have lowest totals of totex and interruption costs. Therefore, UK firms score better when evaluated based on making a better trade-off between costs and quality.

The fact that the efficiency score for Dutch firms increases under the SOTEX model compared to the TOTEX model, confirms the need to include quality into the model's specification. In this specific case, the higher SOTEX scores suggest that although Dutch firms generally have higher costs, a part of these costs is justified from the quality perspective. The way in which quality is incorporated into the model also matters. The difference in efficiency score between the CML and SOTEX models shows that efficiency scores for Dutch firms would be overestimated by an average of seven percent under the CML model. More generally, as previously discussed, the fact that the CML model gives firms that specialise in either quality or costs a high efficiency score does not necessarily imply that these firms provide an optimal quality level at lowest costs. When applied under an integrated price-quality framework, the CML model may thus create perverse incentives for some firms to specialise in quality and for other firms to specialise in costs. Such specialisation is not in line with the idea of achieving an optimal quality level.

An interesting question is whether the low efficiency scores for Dutch firms under the SOTEX models are explained by these firms being less productive or by the fact that their quality level is located further away from the optimum. UK firms have been exposed to incentive regulation much longer than Dutch ones.[6] This makes it plausible that UK firms are more productive and this most likely drives their higher efficiency score. However, the fact that UK firms have less sotex than Dutch firms does not necessarily mean that these firms are providing an optimal level of quality. To illustrate this point, it is helpful to consider Figure 5-5 once again. Let firms A and B correspond to the average Dutch and UK firm. As can be seen, firm A is operating at a lower productivity level but is providing higher quality than firm B. However, both firms A and B are undersupplying quality i.e. a quality increase could well further reduce total social costs. Thus, the fact that Dutch firms provide high quality does not necessarily mean that this quality it too high.

Table 5-4. Change in relative ranking when moving from the TOTEX to the CML model.[5]

	TOTEX	CML	SOTEX
UK-1	1	1	1
UK-2	6	13	6
UK-3	17	18	17
UK-4	4	11	6
UK-5	10	15	14
UK-6	3	1	1
UK-7	9	14	12
UK-8	1	1	1
UK-9	7	10	5
NL-1	16	12	13
NL-2	18	17	19
NL-3	20	20	20
NL-4	14	16	16
NL-5	5	1	4
NL-6	19	19	18
NL-7	12	1	9
NL-8	12	1	10
NL-9	8	1	11
NL-10	11	8	8
NL-11	15	9	15
Average	10	9	10
Average-UK	6	9	7
Average-NL	14	10	13

In summary, although the SOTEX model indicates that there is room for improvement (for Dutch as well as for UK firms), the efficiency scores do not reveal through which mechanisms these improvements can be achieved. This is an important limitation of the SOTEX model. Having said that, it is also important to note that underestimation of the efficiency score is in principle not a problem from the incentive perspective. Given that the translation of the efficiency score into the X-factor takes place in a proper incentive- compatible manner, all firms will still be provided with an incentive to reduce sotex levels. The role of the benchmarking analysis is to identify as good as possible this scope for improvement. Doing so can reduce the extent of rents that firms can earn as a result of beating the regulatory improvement target. However, as the SOTEX model tends to underestimate the true improvement potential, realised improvements may well exceed the potential indicated by the efficiency score and consequently, firms will be able to earn additional profits. This can potentially lead to distributional concerns, in particular if these additional profits become very high. The source of this problem can be traced back to the inferior informational position of the regulator and the fact that the benchmarking analysis provides only an imperfect indication of the firm's true level of performance.

5.4 Integrated Price-Cap Regulation

5.4.1 From Efficiency Score to X-factor

Albeit imperfect, the SOTEX efficiency score reflects the extent by which the firm could operate more efficiently and choose a more optimal quality level. Under the totex approach to price-cap regulation, where the regulator treats opex and capex simultaneously, this efficiency score should be translated into the X-factor. This X-factor would then reflect the regulator's expectations of future improvements in both the productivity as the quality sense. As discussed in chapter three, there are four basic strategies the regulator could follow in translating the efficiency score into an X-factor: Yardstick competition, related caps, isolated caps, and the sliding scale strategy.

Under the isolated caps and sliding scale strategies, the relation between the efficiency score and the X-factor is only indirect. Under the isolated caps strategy, the efficiency score is used to extract information on possible improvements; there is no mechanistic translation of the efficiency score into the X-factor. This is also true under sliding scale, but here the X-factor is adjusted afterwards in the case that profits exceed some predefined band. The isolated caps and sliding scale strategies thus make only indirect use of the information obtained through benchmarking. This is different under the related caps strategy. Here, the efficiency score is converted directly into the X-factor, essentially ignoring any imperfections that may exist in the benchmarking analysis. Similarly, under yardstick competition, the efficiency score directly determines the X-factor although this is done ex post i.e. once the realised improvements have been measured. The role of benchmarking is different in the case of yardstick competition. Here, the general idea is to obtain a measure of change in combined cost-quality productivity that would need to take into account the multi-dimensional nature of the production process (Mikkers and Shestalova 2001). The use of the sotex benchmarking model would thus act as a means to measure such changes in productivity on the basis of multiple output factors.

Irrespective of the chosen price-cap strategy, an integrated price-cap formula would need to consider two basic facts. Firstly, given that the X-factor sets a target for price (productivity) and quality, both these variables should be incorporated into the price-cap formula. That is, interruption costs would somehow need to be internalised into the firm's decision-making process. The second fact is that full internalisation of interruption costs will most likely cause financial sustainability problems as there will be no revenues to cover the additionally imposed costs. In summary, the price formula would need to be designed such that the firm is provided with an incentive to operate in a sotex optimal mode i.e. provides an optimal quality level at least possible cost. At the same time, the regulator would also need to take into consideration the financial sustainability requirement i.e. the firm should earn a reasonable rate-of-return if it operates in accordance with the regulatory targets. To achieve the former objective, the price-cap formula would need to be incentive compatible i.e. generate highest profits in case that the firm reaches an optimal quality level and operates in a productive way. On the other hand, financial sustainability can be assured by internalising not all interruption costs, but only the difference between interruption costs and some predefined target.[7] Assuming no demand variations, the price-cap formula will then take the following form:[8]

$$Rev(t) = (1 - X)^t \cdot \left(c_0 + ic_0\right) - ic(t) \tag{5-13}$$

Here, *Rev* stands for the annually allowed revenue, c_0 and ic_0 stand for the firm's own costs (including a profit element) and the interruption costs, respectively, in the initial year $t=0$. The actual level of interruption costs in a given year is given by $ic(t)$. These two variables together make up total initial sotex and are used as input in the benchmarking analysis. The efficiency score then reflects the extent by which the firm could reduce its initial sotex. This is reflected in the X-factor, which is the annual reduction in sotex levels. Allowed revenues are then obtained by subtracting actually incurred sotex from the annual sotex target. The firm profits will be higher in case that it manages to reduce its sotex levels in excess of the X-factor. The firm's profits (π) can be given by:

$$\pi(t) = Rev(t) - c(t) \tag{5-14}$$

Where *Rev* is the allowed revenue and *c* stands for the firm's private costs. Profits can be reformulated as follows:

$$\pi(t) = (1 - X)^t \cdot \left(c_0 + ic_0\right) - \left(ic(t) + c(t)\right) \tag{5-15}$$

The first part of the equation represents the regulatory target for sotex. Starting from the initial sotex levels at $t=0$, an annual reduction is imposed equal to the X-factor. The second part is the firm's actual level of sotex. If the difference between projected and realised levels of sotex is positive, then the firm earns additional profits. The firm has an incentive to reduce sotex levels faster than anticipated by the X-factor. For example, if most of the savings are made in the first year of the regulatory period, then the firm will earn relatively more profits compared to when savings were achieved in the last year. Also, the firm can increase profits by exceeding the X-factor itself. If the true improvement potential is larger than the regulatory estimate, then the firm may achieve higher profits by beating this estimate. As observed in the previous section, a sotex model tends to underestimate the true improvement potential. Consequently, when the X-factor is based on the efficiency score, the probability of the firm exceeding the X-factor is high. Thus, the firm benefits from the fact that the regulator is unable to conduct a perfect benchmarking analysis.

5.4.2 Evaluation

As was discussed earlier, the possibility to test the statistical significance of efficiency scores is absent due to the deterministic character of DEA. Under the assumption that no modelling errors apply (the given input and output factor definition is correct), the robustness of the sotex benchmarking approach can be evaluated in the light of the three informational problems identified in chapter four.

Firstly, there is the issue of measurement. Any errors in the data may potentially corrupt the efficiency scores. Clearly, the impact of such factors will be more severe under related caps or yardstick competition as in both strategies, the individual efficiency scores and X-factors are interlinked across firms. Secondly, errors in the measurement of demand for quality (as approximated by interruption costs) also influence the outcome of the analysis. Thirdly, with respect to the cost-quality relationship, the benchmarking analysis should recognise that cost and quality levels might vary spatially as well as over time. The proposed integrated price-quality benchmarking approach is now evaluated on the basis of the three informational problems mentioned above.

Cost and Quality Measurement

Integrated price-quality benchmarking is data demanding. The long-term nature of investments makes it important to include multiple years in the benchmarking analysis. As previously discussed, observing only a single year's data may lead to outcomes that are driven by differences in investment timing or accounting practices rather than genuine differences. To make matters worse, including quality makes the benchmark model even more data demanding. Ideally, the regulator would also need to collect quality performance data over a number of years – in addition to the already collected financial data. Experience shows that collecting uniform statistics on quality performance can be problematic. In the UK for example, the regulator observed large variations in the use of definitions and measurement systems across distribution firms (Ofgem 2001).[9] Assuring uniformity in quality measurements is clearly an important requirement to perform integrated price-quality benchmarking. Otherwise, the results of the benchmarking analysis can be corrupted: The efficiency score of one firm can be driven by differences in definitions or even by errors in the data of other firms.

To demonstrate the impact of data (measurement) errors on the benchmarking outcome, a sensitivity analysis can be applied to the empirical analysis from the previous section. The sensitivity of a function to a parameter can be generally defined as the partial derivative of the function with respect to that parameter. This acts as a measure of how much the value of the function will change if the parameter is perturbed. In line with Brown and Ochoa (1998) this can be approximated by actually perturbing the parameter (e.g. by one percent) while keeping all other variables fixed and measuring how much the function changes. For example, if a certain parameter would be increased by one percent and the resulting change in the efficiency score is 0.5 percent, then the sensitivity is equal to 50 percent. Generally, the sensitivity *S* of the efficiency score θ for changes in a variable *v* is defined as:

$$S = \frac{\%\Delta\theta}{\%\Delta v} \tag{5-16}$$

Where *%Δ* stands for the percentage change in the relevant efficiency score or parameter. A change in data for a particular firm may have impact on its own efficiency score as well as on that for other firms. In particular, errors in the data for peer firms i.e. firms with an efficiency score of one will drive efficiency scores for other firms. The reason for this is that these peer firms determine the frontier against which other firms are assessed. A change in the data for these peer firms affects the frontier and consequently potentially changes the efficiency scores for other firms.[10] Data errors are thus particularly relevant for peer firms.

Table 5-5. Sensitivity of efficiency scores under the SOTEX model to a one percent change in the data of peer firms UK-1, UK-6, and UK-8.

	TOTEX			SAIDI			Consumers			Energy		
	UK1	UK6	UK8	UK1	UK6	UK8	UK1	UK6	UK8	UK1	UK6	UK8
UK-1	-94.2%	91.5%	0%	-4.5%	8.2%	0%	0.0%	96.0%	0%	100.5%	8.2%	0%
UK-2	0.0%	32.2%	32%	0.0%	5.4%	4%	0.0%	18.2%	8%	0.0%	5.4%	4%
UK-3	39.3%	30.1%	0%	2.6%	5.2%	0%	-27.5%	18.4%	0%	-13.1%	5.2%	0%
UK-4	39.7%	29.0%	0%	2.1%	4.3%	0%	-29.0%	17.2%	0%	-14.0%	4.3%	0%
UK-5	0.0%	47.5%	2%	0.0%	7.1%	0%	0.0%	27.3%	0%	0.0%	7.1%	0%
UK-6	41.8%	-91.6%	50%	2.0%	-7.0%	7%	-29.9%	-28.9%	15%	-12.9%	-7.0%	7%
UK-7	22.8%	37.6%	0%	1.1%	5.7%	0%	-16.0%	22.8%	0%	-8.0%	5.7%	0%
UK-8	0.0%	39.9%	-88%	0.0%	3.9%	-11%	0.0%	1.9%	-2%	0.0%	3.9%	-11%
UK-9	5.2%	46.1%	0%	0.0%	6.3%	0%	-4.2%	28.3%	0%	-2.1%	6.3%	0%
NL-1	0.0%	0.0%	88%	0.0%	0.0%	11%	0.0%	0.0%	14%	0.0%	0.0%	11%
NL-2	20.7%	37.3%	0%	0.0%	5.5%	0%	-16.6%	23.5%	0%	-8.3%	5.5%	0%
NL-3	0.0%	0.0%	89%	0.0%	0.0%	13%	0.0%	0.0%	0%	0.0%	0.0%	13%
NL-4	7.3%	45.1%	0%	0.0%	6.1%	0%	-6.1%	26.8%	0%	-3.7%	6.1%	0%
NL-5	0.0%	0.0%	88%	0.0%	0.0%	11%	0.0%	0.0%	0%	0.0%	0.0%	11%
NL-6	0.0%	0.0%	88%	0.0%	0.0%	11%	0.0%	0.0%	0%	0.0%	0.0%	11%
NL-7	0.0%	0.0%	89%	0.0%	0.0%	12%	0.0%	0.0%	0%	0.0%	0.0%	12%
NL-8	0.0%	16.5%	61%	0.0%	3.3%	9%	0.0%	8.8%	15%	0.0%	3.3%	9%
NL-9	95.8%	0.0%	0%	4.5%	0.0%	0%	0.0%	0.0%	0%	-99.1%	0.0%	0%
NL-10	0.0%	8.7%	72%	0.0%	1.1%	10%	0.0%	5.5%	17%	0.0%	1.1%	10%
NL-11	95.4%	0.0%	0%	4.8%	0.0%	0%	0.0%	0.0%	0%	-99.0%	0.0%	0%
AVG	**13.7%**	**18.5%**	**28.6%**	**0.6%**	**2.8%**	**3.9%**	**-6.5%**	**13.3%**	**3.4%**	**-8.0%**	**2.8%**	**3.9%**

Table 5-5 shows the sensitivity of efficiency scores of peer firms (respectively UK-1, UK-6, and UK-8) under the SOTEX model with respect to a change in totex, SAIDI, energy distributed, and the number of consumers. As may be observed, a change in the data of peers causes a change in the efficiency scores. This impact is particularly notable for changes in totex levels for firm UK-8. This firm acts as a peer for most Dutch firms and, as can be seen, adjustments in the totex of firm UK-8 have significant impact on the efficiency score for Dutch firms. This may work in two directions: If firm UK-8 would report higher costs than actually incurred, Dutch firms would benefit from this. On the other hand, if the firm would report lower costs, Dutch firms would be provided with an even lower efficiency score.

Demand for Quality

Another area where data errors play a role is in the determination of interruption costs. The sensitivity of efficiency scores with respect to errors in SAIDI corresponds to the sensitivity with respect to interruption costs (as a linear relationship is assumed between interruption costs and SAIDI). In the empirical study presented in this chapter, interruption costs have been approximated based on GNP. This is a rather crude estimation of interruption costs and therefore, a sensitivity analysis as used above may not be the most suitable way to study the impact of errors in the interruption cost measurement method. That is, it may well be that the uncertainty in the GNP proxy exceeds the one percent variation considered in the sensitivity analysis. Therefore, an additional analysis has been performed where the initial estimation of the GNP proxy has been modified in steps by a factor between 0.1 and 10. The results of this analysis are shown in Figure 5-9.

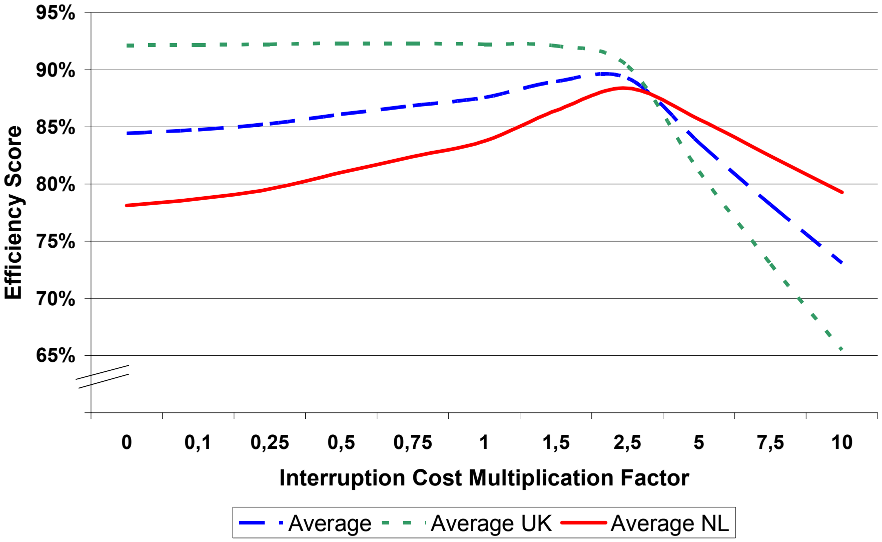


Figure 5-9. Change in efficiency score in percentage points after adjusting the relative weight of interruption costs in the sotex calculation by a multiplication factor.

In line with expectations, efficiency scores for Dutch firms (which generally provide high quality) increase when the relative importance of interruption costs is increased. Similarly, efficiency scores for UK firms show a declining trend. After a multiplication factor of around two, the increase in efficiency score for Dutch firms starts to decline. This can be explained by the fact that a further division takes place between Dutch firms. Firms that provide relatively higher quality than others remain peer while other firms, which provide relatively low levels of quality, start to look less efficient compared to these high-quality firms.

Overall, the impact of an estimation error in interruption costs is moderate within a bandwidth of +/- 0.25 around a multiplication factor of unity. This is because interruption costs are then still only a relatively small share of total social costs. Consequently, a small change in interruption costs has only limited impact on the efficiency score. As the multiplication factor increases, the sensitivity to changes in interruption costs also increases. Note further that in the case that the multiplication factor is zero, the analysis transforms into the TOTEX model.

Spatial Cost-Quality Relationship

The integrated benchmarking model presented in this chapter is based on information measured at the aggregated level. In the measurement of interruption costs, ideally, the regulator should take into account the different quality preferences of consumers as well as other factors that may affect the level of interruption costs. As explained earlier in section 4.3.5, interruption costs are driven by both interruption and consumer characteristics. It is unlikely that the regulator will be able to capture all these factors in the determination of total interruption costs. As a consequence, the regulator's estimation of interruption costs (that enters the benchmarking analysis) will be different from the true level of interruption costs as actually experienced by consumers. In turn, the efficient level of sotex, as identified by the benchmarking analysis, will be biased.

A highly aggregated measure of interruption costs – such as the GNP proxy method used earlier in section three – assumes that all consumers incur equal amounts of interruption costs. In reality, there will be differences across different (classes of) consumers.[11] This can create a bias in the efficiency score due to the fact that the consumer portfolio may differ from one firm to the other. For example, if one firm has a larger proportion of industrial consumers (who generally incur higher interruption costs) then using the same average interruption figure may underestimate the true interruption costs incurred by the consumers. To reduce such potential bias in the efficiency score, the determination of interruption costs could be differentiated by consumer class. This not only requires a measure of interruption costs for each different consumer class, but also information about the quality provided to these consumers. Clearly, this will aggravate the data requirements of the analysis.

Temporal Cost-Quality Relationship

Quality levels will change as a function of costs, but there may be a considerable time lag before the impact of a cost decision becomes visible in terms of a quality change. Thus, at the time the benchmarking analysis is conducted, the impact of quality decisions may not yet be noted. The duration of the time lag between costs and quality is difficult to predict and depends on a number of factors. For example, in the Italian case study (see Annex IV) one can observe a relatively fast increase in quality levels in response to substantial investments in remote control switchgear. On the other hand, it is not possible yet to say what the effects will be of changes in maintenance policy – the full impact on quality is unlikely to become visible until after some years.

Under the integrated price-quality approach as presented here, the X-factor reflects the change in price and quality performance during two successive benchmarking analyses. If the time lag between cost decisions and quality outcomes is short relative to the interval between two benchmarking analysis, then quality trends could be picked up well in time and consequently be reflected in the X-factor. For a given duration of the cost-quality time lag, increasing the frequency of benchmarking and adjustment of the X-factor thus helps to counteract the time lag problem. However, increasing the frequency of benchmarking is typically limited to a period of one year as this is the standard period over which (audited) financial data is available. Furthermore, reducing the duration of the regulatory period diminishes the strength of the incentives as the firm now has a shorter period over which it can retain the benefits of beating the regulatory targets.

5.5 Conclusions

5.5.1 Synthesis

This chapter developed a methodology for integrated price-quality benchmarking under the totex approach. Two basic approaches exist to incorporate quality into a DEA model. On the one side, under the CML model, quality can be defined as a technical output factor. This however has the disadvantage that firms that specialise in cost or quality may be incorrectly classified as efficient. On the other side, the SOTEX model uses total social costs as an input factor. This provides incentives for optimal quality as the firm's efficiency score is based on the extent to which this firm is able to minimise total social costs. The SOTEX model thus provides the right incentives for companies to choose an optimal quality level and doing so in the most productive way as this leads to a higher efficiency score.

The application of the two DEA models to the sample of Dutch and UK firms identified the specialisation problems of the CML model. At the same time, the limitations of the SOTEX model were also illustrated. An important limitation of the SOTEX model is that the efficiency score does not differentiate between improvement potential in productivity or quality. Also, an efficiency score of one does not necessarily mean that the firm is providing the true optimal quality level nor does it mean that the firm is fully productive. The efficiency score thus may well underestimate the true improvement potential in both the productivity and the cost sense. Moreover, due to the fact that consumer preferences for quality differ, the use of aggregated interruption cost data also creates bias in the efficiency scores. It is not likely that the regulator can identify the true level of interruption costs, although the use of interruption cost assessments disaggregated by consumer class can help to reduce this bias.

The deterministic nature of DEA does not allow any statistical checking of the obtained efficiency scores. This is particularly a problem as no provisions can be made to reduce the impact of data errors on the benchmarking outcome. Assuring high quality data is thus an essential condition for carrying out an effective benchmarking analysis. This is particularly true when the regulator opts for a price-cap strategy based on related caps or yardstick competition. These strategies impose a direct relation between the efficiency score and the X-factor and thus are more sensitive to data errors. Here, errors in the data of one firm can potentially affect the X-factor of other firms.

The sotex benchmark tends to underestimate the firm's true improvement potential in the realms of cost and quality but this, in principle, is not a problem from the incentive point of view. There is still an incentive to reduce sotex beyond the regulator's targets, as this is associated with additional profits. However, if these extra profits become too large, then distributional problems may emerge. Of a more severe nature is the fact that the integrated price-quality benchmarking may not be effective if there is a long time lag between cost decisions and quality outcomes. In principle, the dynamic relation between costs and quality is captured by a sotex model as any change in costs or quality is eventually reflected in the firm's X-factor. However, if the firm discounts future profits at a too high rate, then it could strategically exploit the time lag between costs and quality. That is, the firm may aim at short-term profit maximisation and simply disregard the adverse effects on profitability on the longer term. Recent financial scandals such as the Enron case suggest that the risk of the firm (or more specifically, the firm's managers) adopting a short-term view is indeed not unrealistic.

Since is not likely that the benchmark analysis will pick up adverse reductions in quality well in time, it may be worthwhile setting in place additional monitoring systems. The regulator could for example closely monitor developments in quality levels as well as in the firms' spending behaviour. The question then is what happens once the regulator finds that the firm is not investing sufficiently in the regulator's view. Imposing minimum spending requirements would clearly not be in line with the price-cap philosophy. Indeed, the basic idea under the totex approach is that the regulator sets in place targets based on some assessment; thereafter the firm is free to decide on its own spending levels without any further interference by the regulator. Requiring minimum spending by the firm would not be in line with this line of thinking.

Having observed some limitations of integrated price-quality benchmarking under totex, it is also worthwhile to point out its value. Given that data quality is reasonable, the benchmarking analysis does provide valuable (albeit imperfect) information to the regulator. The question is not whether benchmarking should be applied or not, but rather how the results of the benchmarking analysis could most appropriately be used. In the face of data uncertainty, a strategy based on isolated caps or sliding scale may be more appropriate than a related caps or yardstick competition strategy. Here, there remains room for regulatory discretion in translating the efficiency score into the X-factor. This also creates scope for imposing additional constraints in order to deal with the time lag problem. Requiring minimum spending is one example of this. Effectively then, the regulator moves away from a pure totex approach towards an approach that is more similar to building blocks. Taking this further, the regulator could prescribe a certain level of spending – one that is believed to result in an optimal quality level and reflects an efficient mode of operation by the firm. For this approach to be effective however, the regulator would need to obtain information about the cost-quality relation i.e. predict how the firm's cost decisions will affect (future) quality levels. The better the regulator is able to assess this relationship, the more accurate can he set an appropriate spending target.

5.5.2 The Way Forward

This chapter dealt with integrated price-quality regulation under the totex approach. In the next chapter, an alternative methodology for integrated regulation will be developed in which the regulator explicitly prescribes cost decisions based on an assessment of the cost-quality relationship. This methodology fits under the building blocks approach and is based on the application of integrated price-quality benchmarking for individual network investments.

Notes

[1] An overview of DEA is in Seiford and Thrall (1990). Two basic DEA models to which most applied papers still refer are those by Charnes, Cooper, and Rhodes (1978) for constant returns to scale (CRS) DEA, and by Banker, Charnes, and Cooper (1984) for variable returns to scale (VRS) DEA.

[2] The term allocative efficiency used here has a different meaning than the one used in chapter two.

[3] The choice of input and output factors tend to vary. A discussion of the choice of input and output factors in DEA is provided by Rossi and Ruzzier (2000).

[4] Costs and quality are not necessarily linearly related. This does not change the basic observations here.

[5] All firms with an efficiency score of one are assigned a ranking of one.

[6] Incentive regulation for UK firms started in 1990 while in the Netherlands, this only started in 2001.

[7] The financial sustainability requirement implies that the firm earns a predefined rate-of-return only if it manages to operate in line with the regulator's improvement potential estimations as reflected in the X-factor.

[8] In reality, demand will change over time and consequently impact costs. This can be taken into account by including additional demand drivers in the price-cap formula (Green Rodriguez-Pardina 1999).

[9] Assuring uniformity in quality measurement is particularly problematic due to the fact that not measuring interruptions can actually make the firm seem to provide higher quality.

---

[10] It may also be possible that a firm no longer is located on the frontier after a data modification, and conversely.

[11] See also Table 4-4.

6

Network Simulation Tool

6.1 Introduction

6.1.1 Background

This chapter deals with the issue of integrating quality into the price-cap system when use is made of the building blocks approach. Under the building blocks approach, the firm proposes certain investments to the regulator who in turn needs to determine whether these may be charged to the consumers. Approved investments are included into the RAB and prices reflect an allowance for their depreciation as well as a rate-of-return on these investments. Under an approach where the firm's returns depend on the level of investments allowed into the RAB, there is a risk of the overcapitalisation or Averch-Johnson effect. In this particular case, the Averch-Johnson effect takes the form of inflated investment projections. In the spirit of price-cap regulation, the firm has an incentive to reduce investment levels if it can retain (part of) the difference between the forecasted and actual investment level. This creates a natural tendency for the firm to inflate its investment proposals. That is, the firm increases its profits by inflating the size of the RAB but not necessarily undertaking these projected investments.

An important regulatory task under building blocks is to ensure that the firm invests at an appropriate level. This level should also be reflected in the determination of the RAB. At the same time, the regulator also needs to ensure that the investments undertaken by the firm provide consumers with a desirable quality level. However, due to his inferior informational position, it is difficult for the regulator to evaluate whether a given investment proposal is effective. That is, to assess whether the investment is conducted in a cost-efficient manner and whether it results in a socially optimal level of quality.

As was discussed earlier in this thesis, benchmarking is an important regulatory tool for dealing with the informational problem. Models for opex benchmarking are nowadays widely applied while the previous chapter developed a retrospective price-quality benchmarking model under the totex framework. The application of traditional benchmarking models (such as DEA) in the area of investment evaluation is limited, however. The heterogeneous character of investments makes it difficult to construct a comparable and sufficiently large data sample. On the one side, each investment is characterised by unique demand and supply conditions, which leads to substantial differences in both cost and quality performance. This makes it difficult to compare individual investments in a straightforward way. Although structural differences could possibly be captured in the benchmarking model, this would require an even larger sample size in order to maintain the discriminative power of the analysis. On the other side, firms have some flexibility in the timing of the investment. This not only reduces the sample size further but also creates some scope for strategic allocation of investments over time.

The problem of evaluating investment performance is further complicated by the observation that it is important to consider the integrated costs and quality performance of the investment. Investments conducted at low costs may not necessarily be effective as they may provide consumers with insufficient quality. Similarly, very expensive investments may be associated with an oversupply of quality. The regulatory challenge lies in the integrated evaluation of both price and quality performance of each investment and in assuring an optimal balance between these two variables. Thus, investment appraisal is one of the areas where the regulatory informational disadvantage is most evident.

At the same time, capital costs generally form a substantial part of the firm’s total costs and investment decisions have a significant impact on the network’s quality. Ensuring that investments are undertaken at least costs and deliver a socially optimal quality level can generate significant benefits to society. A tool that enables the regulator to effectively measure the performance of investment proposals with respect to their effects on both the price and quality level of the service provided is therefore an important regulatory asset.

The objective of this chapter is to develop a benchmarking methodology that can be used for integrated price-quality assessments of network investments. This novel methodology is aimed specifically at evaluating investments in new distribution networks. The methodology is implemented in the form of a software tool programmed under Matlab – the Network Simulation Tool (NST). The NST is based on the idea of comparing the performance of the firm’s investment proposal to that of a large number of artificially constructed alternatives. These alternatives are generated through simulation and represent possible solutions that the firm might have considered instead of the one being actually proposed. The performance of an investment is measured in terms of its total social costs (sotex), which is defined as the sum of network costs and interruption cost. Comparing the performance of the actually proposed investment and the artificially constructed alternatives provides information that can be used by the regulator to determine the RAB and subsequently set the appropriate X-factor under the price-cap scheme.

6.1.2 Chapter Outline

This chapter is structured as follows. Section two starts with a general description of the NST and reviews the NST in the light of existing network models. Section three presents a more detailed description of the NST. Section four applies the NST to a case study of a fictive Greenfield distribution network.

6.2 General Design

6.2.1 Alternative Network Construction

Rather than comparing a given network investment proposal to other actual investments, the NST generates the possible alternative solutions for the investment under scrutiny, and evaluates whether these alternative solutions perform better than the proposed one. Starting from a given Greenfield supply area, the basic idea of the NST is to construct a large number of network designs and choose the most effective one from these. In constructing the network alternatives, as much as possible variation in network characteristics is allowed, thus increasing the probability of arriving at the true optimal network. If one could enumerate and evaluate all theoretically possible networks for a given Greenfield area, the best alternative could simply be identified and subsequently used as the benchmark. However, the total number of alternatives that need to be analysed tends to increase exponentially as a function of the size of the network. Consider for example a simple network consisting of 10 connections (e.g. cables) the routing of which is fixed and where the only decision variable is the type of cable. If there are five possible cable types to choose from, the total number of networks would already be equal to 5^{10}=9.7 million. If, in this simple example, a computation time of 10 microseconds per network is assumed, the total computation time for the analysis would already amount to more than one day. If a more realistic network is considered, taking into account, for example, additional connections, routing possibilities, network protection schemes, variations in equipment types, then the number of alternative networks would increase at an exponential rate. Even for relatively small networks, the total number of alternatives to be evaluated would quickly grow very large and lead to unpractical computation times.

An approach where one enumerates and evaluates all possible networks is not practical due to the constraints in computational power. It is important, however, to recognise that it is not necessary to consider all possible networks. The number of networks – and therefore computation time – can be substantially reduced by adopting a smart network selection strategy. From the full set of all possible networks, only a small selection will be truly suitable for consideration in practice. Most networks will either be too expensive or provide a too low level of quality. Only a relative small portion of all possible networks will offer a proper balance between price and quality i.e. will be effective in the integrated price-quality sense. By limiting the analysis to these networks only, the number of network alternatives to be considered can be significantly reduced and hence computation time reduced accordingly.

Particularly, a faster convergence towards the theoretically optimal network can be realised by choosing an iterative approach where one quickly narrows the scope of the analysis to network types that exhibit superior performance. For this purpose, a genetic algorithm as shown in Figure 6-1 can be applied. The first step of the algorithm consists of constructing an initial set of networks that could be used for a given newly to be supplied Greenfield area. Each initial network has a given set of basic features, which may relate to aspects such as the number of feeders, the choice for a radial or meshed design, the type and quality of assets, the protection system, etc. In principle, as much flexibility as possible in the basic features of the networks is allowed in this stage of the process.

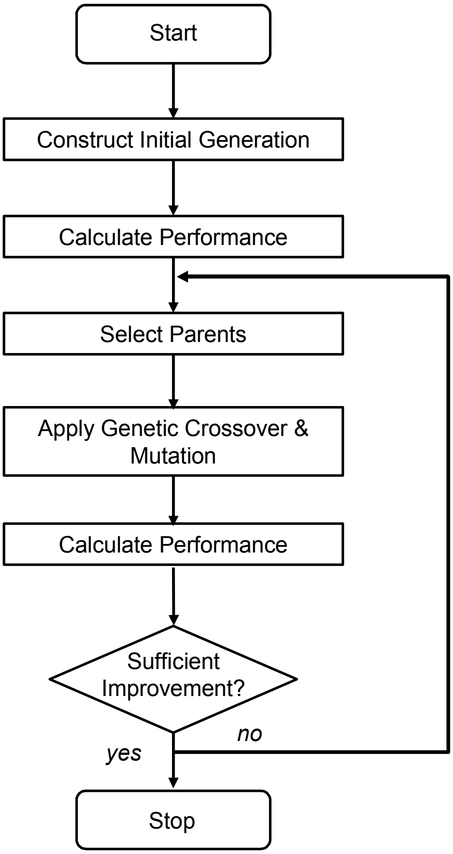


Figure 6-1. Basic steps of the NST.

These initial networks are denoted as the "parents". In the next round, a new set of networks - "children" - is constructed through the combination of parental networks. This new generation of networks inherits certain key features from the parent networks. However, not all initial networks are used in the process of creating the next generation of networks. Each initial network is assigned a certain probability of being considered a parent network. Networks that exhibit superior performance – that is, have low levels of sotex – have a higher probability of being selected as parents than networks which perform poorly.

The construction of new networks takes place by using a genetic crossover technique where the basic features of the two parental networks are randomly combined in order to create a new network. Furthermore, for maintaining genetic diversity, a mutation operator is applied. Here, a certain probability is introduced for one or more basic features of the new network to change its value beyond the constraints of the genetic crossover.

Calculating the sotex performance of each network consists of three steps. Firstly, a reliability analysis of the network is conducted. From this, interruption costs experienced by consumers can be obtained. Secondly, the cost of the network itself is determined on the basis of the quantity and price of the network assets as well as the cost of losses. Finally, adding up interruption and network costs provides the measure of total social costs or sotex. The process of creating subsequent generations of networks is repeated for a number of times until no substantial further improvement in the networks' performance occurs.

6.2.2 Experiences with Network Models

The idea of using a model to construct artificial networks for benchmarking is not new. Network models for electricity distribution are, among others, used by regulators in Sweden, Chile, and Spain.. These models are now briefly discussed.

The Swedish regulator has developed a network model known as the Network Performance Assessment Model (NPAM).[1] The NPAM constructs a single fictive network based on several input data regarding the geographical location of the loads, electricity consumption, and connections to other networks. The fictive network consists of multiple voltage levels. Starting with the low voltage network, consumers are grouped and connected to a transformer. This grouping is performed on the basis of different conditions such as the expected voltage drop and the consumers' distance to the transformer. Constructing the network starts by putting a transformer in the so-called electricity gravity centre for the given consumer group. Consumers belonging to this group are subsequently connected to this transformer. This process starts with the connection to the transformer of the consumer that is closest to it. Then, each subsequent consumer is connected to the transformer or to an already connected consumer, whichever is closest. This process is repeated for all groups until all the consumers within each group have been connected to a transformer. The same principle is then used to connect the different transformers to a transformer positioned in the electricity gravity centre of the next voltage level. Eventually, through this cascading approach, a fictive network is constructed which forms the basis for calculating network performance.

The calculation of the network performance, which is the main output of the NPAM, takes place in two steps. The first step provides the network cost. This consists of opex, capex, and network losses. Opex is estimated as the sum of administration, operation and maintenance costs. The administrative cost is assumed proportional to the number of clients and covers the cost of meter depreciation, meter reading, invoicing, and processing. Other operational and maintenance costs are assumed proportional to the asset value. Capex is estimated on the basis of the purchase price of the assets in the fictive network and includes an allowance for depreciation as well as a rate-of-return. Network losses are estimated for each connection point using functions that depend on the density of the fictive network and are valued at a standard price.

The second step in calculating the network performance consists of calculating the quality cost. This is done by considering the interruption cost incurred by consumers on the basis of the reliability level they experience. The cost of quality is thus not derived from the quality performance of the fictive network, but from the actual quality level that consumers experience. The reason for this, according to Gammelgard and Larsson (2003), is that quality is regulated separately outside the NPAM. In determining the quality cost, a distinction is made between consumers based on the density of the network. The principle followed is that interruptions are more expensive for consumers located in dense networks (e.g. urban areas) and that unannounced interruptions are more expensive than announced ones. Based on this, the total interruption cost for the network is calculated. The network total performance is then defined in terms of the sum of the network cost and the interruption cost. This is the benchmark against which firms are compared and prices are envisaged to be set by the regulator.

In Chile, network models are an important aspect of the regulatory process. Here, an ideal firm is constructed based on the actual demand and the expected load growth. Rudnick and Raineri (1997) and Rudnick and Donoso (2000) provide descriptions of the Chilean network model. Starting from the existing grid configuration and assets, a model is applied to optimise the maintenance, operations, and management of the firm. The model takes into account fixed costs such as administration, invoicing and user service expenses as well as variable costs, which include network losses, investments, operational costs, and maintenance costs. In order to maintain comparability between firms, different network zones (high density, urban, semi rural and rural) are identified where each zone represents an area of homogeneous technical and economic conditions. For each firm, the network model determines the cost that would be incurred by an efficient firm supplying electricity to a mixture of zones corresponding to the actual firm. As far as known, quality is not directly considered in the optimisation process.

Both the Chilean regulator and the distribution firms perform optimisation studies using different sets of models. The cost corresponding to an efficient firm is defined as the weighted average of the regulator's estimation of the optimal cost and that of the firm where the weights are set at 2/3 and 1/3 respectively. The results of these studies form the basis for determining the firm's income. The models used for the analysis are, however, not public and the technical details of the calculation of the cost incurred by the model firm are not disclosed due to the highly detailed nature of the analysis. Nevertheless, the fact that different studies are conducted and that these studies tend to generate different results indicates that apparently, there is not one common optimisation methodology; different methodologies are employed. For example, Rudnick and Raineri (1997) report that some firms claimed a level of optimal costs that was twice as high as the regulatory estimate. This is not surprising as one may assume firms to have an incentive to inflate their estimation of costs in order to receive higher revenues.

Under the Spanish regulatory framework, use is made of a network model known as BULNES (Sumicsid 2003a). The BULNES model determines the optimal network cost for each distribution firm based on information regarding the geographical position of consumers, demand levels, and the connecting nodes to the transmission network. Using this information, optimisation algorithms are applied to determine the location of transformers, routing of lines and cables, etc. The cost of this optimal network is then derived using assumptions on the price of the assets and the unit cost of preventive maintenance, corrective maintenance, and operations. Furthermore, prices are adjusted to the particular case of each firm to reflect different conditions affecting the operation of the system that are related to ice, salt, precipitation, altitude, and rights of way. BULNES does not consider interruption costs incurred by consumers. Rather, reliability is considered in the form of the cost of carrying out corrective maintenance based on the expected number of interruptions and the associated repair costs.

The information provided by BULNES is used by the regulator to determine the efficient cost of distribution corresponding to each firm and to allocate the total revenues to be received by the distribution firms among them. Revenues are allocated proportionately to the efficient cost for each firm. The total income level for the industry is predetermined on the basis of agreements between the regulator and the industry. The purpose of the BULNES model is thus to allocate this income amongst the different firms rather than determine the absolute level of the sector's income.

Another network model developed in Spain is proposed by Peco and Gómez (2000). This model is similar to the BULNES model except for the fact that it explicitly considers quality. The model constructs an optimal distribution network based on the exact geographical location of the consumers and their associated demand. For each service area, the model constructs the optimal network and calculates the associated network costs. This optimal network links loads with generation sources assuming there are no equipment failures; its configuration is optimised as far as the cost of investments and losses are concerned. The optimisation process is subject to operational constraints on the feeder capacity and the magnitude of voltage drops. It makes use of heuristic algorithms that design a radial network linking sources and consumers. This ideal model network is built from scratch minimising both the investment and the cost of network losses. In doing so, possible constraints imposed by the geographical location of nodes (e.g. right-of-way) are also taken into account.

Once the optimal radial network has been obtained, a second optimisation process decides the number, location, and size of feeder reinforcements and new feeder sections that create alternative routes of supply to the loads. This process also determines the optimal level of investment in switching and protective devices. This second optimisation stage is aimed at minimising the total social cost associated with the network taking into account the cost of interruptions. During this process, network reinforcements are included and discarded until the sum of network and interruption costs cannot be further reduced, i.e. the least cost network from a social point of view has been obtained.

6.2.3 Network Simulation Tool versus Traditional Models

There are some important differences between the NST and the traditional network models discussed above. The outcome of traditional models is a single network that results from an optimisation process. In contrast, the NST is not based on optimisation but rather on the principle of simulating a large number of networks, out of which the most optimal one can be selected. The advantage of this approach is that it recognises the limitations connected to the use of optimisation algorithms and leaves – in principle – full flexibility in the choice of network characteristics. In literature, a number of algorithms for optimisation of the design of electricity distribution networks have been presented.[2] There are important differences between these algorithms as they are based on different assumptions regarding the desired network topology. The choice of an optimisation algorithm a priori constraints the outcome of the analysis. For example, certain algorithms only consider networks with a radial design and thus automatically exclude meshed networks as a viable outcome. However, it may well be that in the particular case a meshed network performs better.

Another and perhaps more important feature of the NST is that it considers costs and quality in a fully integrated fashion. Networks are evaluated both in terms of their cost and quality performance. Traditional models either ignore quality or consider quality only in a second stage. For example, the Swedish model does not consider the quality performance of the optimised network but assesses the network performance on the basis of the number and duration of the actual interruptions experienced by consumers. The model by Peco and Gómez (2000) is the one that comes closest to being an integrated price-quality approach. Here, quality enters the analysis in the second stage by adjusting the network that was obtained in the first stage of the optimisation process. This is an important limitation as the outcome of the second stage of the process, where quality performance is considered, is conditioned by the outcome of the first stage. For example, a certain network may initially be regarded as expensive when evaluated only considering the cost side but may well turn out to provide a better price-quality trade-off than other networks. This network will, however, not be considered in the second-stage of the analysis if it has already been discarded in the first stage.

The unique feature of the NST on the other hand, is that it considers price and quality performance simultaneously. The performance of a network is defined in terms of the total social cost and, based on this measure, subsequent networks are constructed and evaluated. This allows the search process to consider all types of networks as long as their overall performance, as measured by the total social cost, is satisfactory. Table 6-1 summarises the main differences that exist between the NST and traditional network models.

Table 6-1. Summary of differences between traditional models and the NST.

Model	Basic approach	Treatment of Quality
Sweden (NPAM)	Cascading algorithm	No
Chile	Unknown	No
Spain (BULNES)	Optimisation algorithms	Indirectly
Spain (Peco and Gómez 2000)	Optimisation algorithms	Second stage optimisation
NST	Simulation	Integrated with price

6.3 Detailed Design

6.3.1 Construction of Initial Networks

The NST considers only the distribution network; transmission and low voltage networks as well as production facilities are excluded from the analysis. The assumption is that the network consists of one main feeding point (represented by a HV/MV substation) and a variable number of MV substations. Furthermore, loads are assumed to be represented by a number of MV/LV transformer stations. Each MV substation is supplied by the HV/MV substation through one or two primary feeders. The MV/LV transformer stations are in turn supplied by the MV substations via the secondary feeders. The location of the MV/LV transformers as well as information about demand characteristics are input data to the NST. Figure 6-2 provides a schematic representation of the type of systems the NST can be applied to.

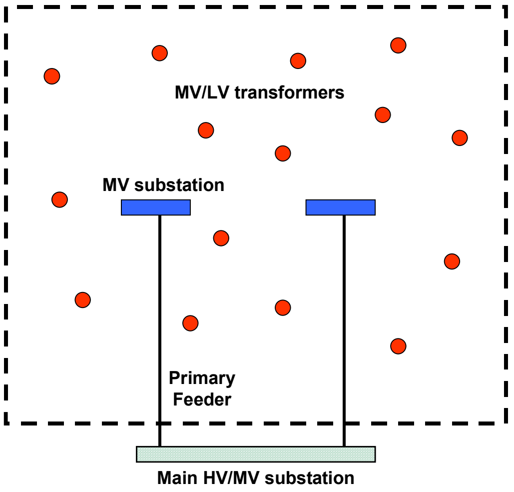


Figure 6-2. Schematic overview of the scope of the NST.

The construction of networks essentially takes the form of combining the basic features of two parent networks. These features can be represented in terms of a string of bits where each bit contains information about the network characteristics.[3] Generally speaking, two types of bits can be identified. Firstly, a number of bits that represent key characteristics of the network. Secondly, a variable number of bits that contain information about the connections in the network. These in turn consist of information regarding the existence of an electrical connection between two nodes as well as information about the type of connection assumed (e.g. cable or line and the capacity, reliability and resistance of that cable or line). In the construction of new networks, the strings of bits of the two parental networks are randomly combined into a single one. The combination of bits – that is, the construction of the new network – takes into account a number of constraints. To begin with, there is the adequacy constraint, which requires that all loads are supplied in the base situation, i.e. when no faults have occurred in the network. Furthermore, there are some physical constraints that need to be taken into account: The resulting current flows in the network will dictate the minimum capacity of each feeder and consequently limit the choice of the cable or line type. An overview of the possibilities regarding the key characteristics of networks is shown in Table 6-2. A detailed description of these options is contained in Annex VIII.

Table 6-2. Overview of choices with respect to the key characteristics of the network.

Key characteristic	Allowed Range
Number of MV substations	1, 2, 4
Number of primary feeders per MV substation	Single or Double
Number of secondary feeders per MV substation	1-31
Basic network topology	Radial or Meshed
Protection for primary feeders	Yes / No
Protection for secondary feeders	Yes / No

By combining the key characteristics of the network as well as varying the routing within the network, different networks can be constructed. Each network will perform differently with respect to costs and quality. Generally, adding more redundancy to the network increases quality but also comes at an additional cost. For example, doubling the number of primary feeders leads to an increase in quality since a failure in one of the feeders no longer leads to an interruption (assuming that each feeder has sufficient capacity to supply all load). Similarly, increasing the number of secondary feeders leads to relatively shorter feeders. This, on the one side, reduces the impact of an interruption whereas, on the other side, shorter feeders also lead to a lower probability of an interruption taking place in that feeder. Conversely, installing less protection increases the impact of the interruption while designing the network in a radial way results in longer interruptions due to the lack of possibility to reroute power flows in the case of interruptions.

Clearly, there are a large number of possibilities to choose from, but not every quality improving or reducing measure will be effective from the social point of view. The underlying idea of the NST is to consider the impact of these different choices regarding the design of the network and, on the basis of this information, enables the regulator to more effectively evaluate the performance of a given investment proposal.

6.3.2 Genetic Crossover and Mutation

For selecting networks that will act as parents for the subsequent generation of networks, a mix of the elitist class and tournament selection methods is used.[4] Under the elite class method, networks are sorted in order of their sotex performance and a selection is made of the best performers, which become part of the so-called elite. For example, the elite can be defined as the 10 percent best performing networks. This selection method makes sure that alternatives with relatively high performance (i.e. low levels of sotex) are always maintained as parents.

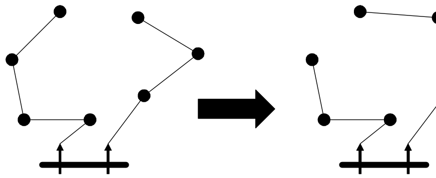
Table 6-3. Simplified example of genetic crossover to produce two children.

Key characteristic	Parent I	Parent II	Child A	Child B
Number of MV substations	1	2	1	2
Number of primary feeders per MV substation	1	2	2	1
Number of secondary feeders per MV substation	2	3	3	2
Basic network topology	Radial	Meshed	Radial	Meshed
Protection for primary feeders				
• Primary Feeder 1	Yes	No	Yes	No
• Primary Feeder 2		Yes	Yes	
Protection for secondary feeders	Yes	No	No	Yes

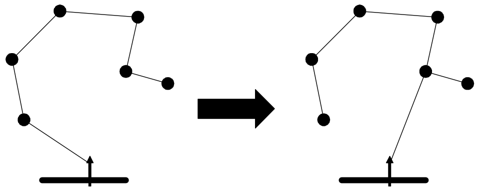
The elitist selection method, however, has the disadvantage that genetic diversity is reduced over time. To counteract this potential problem, an additional set of parents is selected through the tournament selection method. Here, a random selection of two networks is made and the one with least sotex is classified as a parent. The probability for better performing alternatives to be selected is thus higher while at the same time, poor performers still maintain a chance to be selected and transfer their genetic information to subsequent generations. Note that a given network may be selected as parent more than once. For example, an elite network is not excluded from participation in the tournament and thus may be selected under both the elitist class and tournament selection methods. Furthermore, a given network can also be selected multiple times under the tournament method.

Once parent networks have been chosen, a genetic crossover is applied to create a new generation of networks. Each combination of two parents produces two children. The binary string of the new network is determined by a random selection between the binary strings of the two parent networks. The new network thus inherits the combined characteristics of the two parent networks. An example of this procedure is shown in Table 6-3.

In addition to genetic crossover, the newly constructed network is exposed to mutation. The importance of mutation lies in the fact that it provides for a constant supply of fresh networks with increasing genetic diversity. In this case, two types of mutation are applied. Firstly, a small probability is introduced for any of the network's main characteristics to randomly change its value. Secondly, mutation is applied to the network's routing scheme through branch replacement techniques. Here, the network's routing scheme is randomly reconfigured. Two types of branch replacements techniques are used. Changes of type I modify the network's routing in such a way that a given load point is supplied through a different feeder. Here, only a single feeder segment is adjusted at one time. Under branch replacements of type II on the other hand, two feeder segments are modified at the same time. Each branch replacement takes into account the constraint that supply to all loads is maintained. Figure 6-3 provides an example of each of the two respective branch replacement techniques. An overview of the algorithms used for conducting the branch replacements is presented in Annex IX.



(a) Branch replacement of type I.



(b) Branch replacement of type II.

Figure 6-3: Example of type I and type II branch replacement techniques.

6.3.3 Reliability Analysis

The reliability analysis provides information about the frequency and duration of interruptions experienced by each consumer.[5] This information can then be used to calculate the level of interruption costs which is one of the components of the network alternative's sotex performance. For an individual consumer *i*, annual interruption costs *ic* can be defined as follows:

$$ic_i = ENS_i \cdot ICENS_i \tag{6-1}$$

And total interruption costs *ic* can be derived from:

$$ic = \sum_i ENS_i \cdot ICENS_i \tag{6-2}$$

Here, *ENS* stands for Energy Not Supplied which is the annual amount of energy not supplied to consumers. *ICENS* is the interruption costs incurred per kWh of *ENS* and is an input parameter. The annual *ENS* per consumer in turn can be approximated by:

$$ENS_i = f_i \cdot d_i \cdot PD_i \cdot LF_i \tag{6-3}$$

The frequency f and duration d of interruptions affecting consumer i are the output of the reliability analysis. The parameters *PD* and *LF* are input parameters and denote the peak load and the load factor, respectively, for consumer *i*.

For calculating the reliability performance of the network, a computer program named Distrel has been integrated into the NST.[6] Distrel is based on a probabilistic approach to reliability analysis that has as a starting point the failure of network components. This failure behaviour can be modelled as a Markov process. Here, a given network component is thought to be in either one of two states namely the UP-state (the component is functioning normally) or the DOWN-state (the component has failed). Moving from the UP-state to the DOWN-state means that the component has failed and conversely, moving from the DOWN-state to the UP-state implies that the component has been repaired. Moving from one state to the other occurs with a certain probability and can be expressed in terms of failure rate and repair rate.

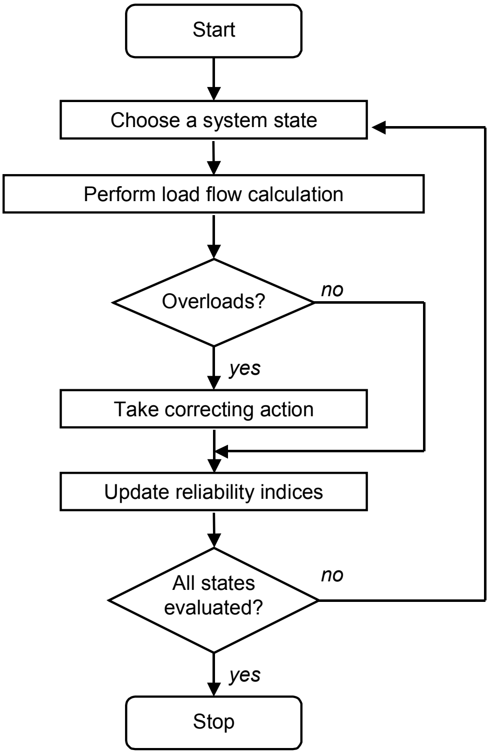


Figure 6-4. Basic steps involved in conducting a reliability analysis under Distrel.

The failure rate reflects the probability of moving from the UP to the DOWN state and can be determined empirically by observing the actual performance of that component during a long period of time. If F_k is the number of failures of a component k observed during a predefined period (typically one year) and $t_{u,i}$ is the time a component k is in the UP-state, the failure rate can be defined as:

$$\lambda_k = \frac{F_k}{\sum_{i=1}^{F_k} t_{u,i}} \tag{6-4}$$

Once a component has failed, it will not remain in the DOWN-state but will be brought back to the UP-state by repairing it (or replacing the component by a new one). The repair rate represents the probability of a given component to move from the DOWN to the UP state. From this, the repair time can be derived i.e. the average time taken to repair the component. If $t_{d,i}$ is the time that the component *k* resides in the DOWN-state, then its repair time can be defined as:

$$\mu_k = \frac{\sum_{i=1}^{F_k} t_{d,i}}{F_k} \tag{6-5}$$

In the case of meshed networks, there is usually the possibility to restore supply by means of rerouting power via alternative feeders. The faulted component can then be electrically isolated and repaired while supply is already restored. The time the component resides in the down state is then reduced. Generally, it is helpful to make a distinction between repair time and switching time where the latter replaces the former in case of rerouting possibilities.

The failure rate will generally not be constant but will vary with the age of the component. Typically, the failure rate behaves in line with the so-called bathtub curve with the probability of failure being higher during the early and later years of the component lifetime (Klaassen et al. 1988). In the intermediate period, which is the largest part of the component lifetime, the failure rate is more or less constant. For practical purposes, this constant failure rate is used in the application of the reliability analysis. Similarly, the repair or switching time may vary as a function of different factors but for the purpose of the simplified analysis is here assumed constant.

The failure and repair characteristics of an individual component represent the probabilities of moving from one state to the other. By considering all network components simultaneously, a system state can be defined. A system state is defined as a combination of the states of all individual components and consists of components being either in the DOWN or UP state. Each system state may or may not involve interrupted consumers. If the number of system states is equal to *N*, then probability *Pr* for a certain system state occurring can be derived by solving the following set of equations (Meeuwsen 1998a):

$$Pr(l) \cdot \sum_{m \in N} \tau_{lm} = \sum_{m \in N} Pr(m) \cdot \tau_{ml}, \quad l = 1,2,3,\ldots\mathrm{N} \tag{6-6}$$

Here, *Pr* is the probability of a certain state to occur and τ_{lm} is the transition rate from state *l* to state *m*, which follows from the failure and repair times of the different components. The average duration *Du* of residing in a certain state is equal to:

$$Du(l) = \frac{1}{\sum_{m} \tau_{lm}} \tag{6-7}$$

And the frequency of being in that state is:

$$Fr(l) = \frac{Pr(l)}{Du(l)} \tag{6-8}$$

From this information, the cumulated interruption frequency and duration indices (respectively *f* and *d*) for each consumer can be obtained. The steps involved in carrying out this analysis are shown in Figure 6-4. Firstly, a certain system state is selected and it is assessed whether any consumers would be interrupted in the given system state. This can be done by conducting a load flow analysis. If a certain component (or combination of components) has failed, then this may or may not lead to any consumers experiencing service interruptions. Furthermore, other components may become overloaded and thus corrective actions will need to be applied. This may also lead to additional consumers being interrupted as the system's protection will automatically disconnect these overloaded components. The duration of the interruption in turn depends on the possibilities to restore power through switching actions or the time it takes to repair the faulted components. This process is repeated for all possible states in order to eventually determine the interruption frequency and duration for each consumer.

It is clear that, when the number of components increases, the number of system states grows exponentially and consequently, the computation time required to perform the reliability analysis may become very long. In order to reduce the computation time, here, only first order states are considered. That is, for a given system state, only one component is assumed to fail at any given point in time. By excluding states where multiple component faults occur, the number of system states to be analysed can be reduced substantially. At the same time, this does not have a large impact on the accuracy of the results obtained, as the probability of two or more components failing at the same time is relatively low.[7]

6.3.4 Cost Analysis

The second part of the sotex evaluation consists of computing the costs of the network itself. Network costs can be divided into investment costs and the costs of losses occurring in network cables and lines. Investment cost comprises the cost of assets belonging to:

- Substation buildings;
- feeder bays including breakers and disconnectors;
- feeder protection systems;
- primary and secondary feeders (km cable or lines).

The number of substations, feeders, disconnectors, and protection systems directly follow from the network's configuration. For determining the length of the feeders, the distance between each pair of connected nodes (being either a MV substation or a MV/LV transformer station) needs to be considered. Feeders generally follow roads, streets or property boundaries which means that the feeder's segments are built along a rectangular pattern of roads and streets rather than the shortest path between two given points (x_1,y_1) and (x_2,y_2). The Euclidean Distance consequently tends to underestimate the length of the feeder (Willis 1997, p. 291). Therefore, the Lebesque or "taxicab travel" distance ($|x_1-x_2|+|y_1-y_2|$) is used since it is a more reliable estimate of the actual length of the feeder.

The prices and volume of assets determine the total investment cost (*capex*) that will need to be recovered during the lifetime of the assets. Capex can be divided into two parts, namely depreciation and return. The depreciation costs refer to the initial purchase price of the asset while the return is associated with the capital costs of the asset. The costs of capital consist of both the costs of debt (interest) and the costs of equity (dividends). For comparison purposes, it is helpful to express capex in terms of an annuity. If a lifetime of *n* years is assumed, then the annual levelised *capex* for a given asset of type *a* can be given by Willis (1997, p. 209):

$$capex_a = \frac{ror \cdot (1+ror)^n}{(1+ror)^n - 1} \cdot p_a \cdot q_a \tag{6-9}$$

Where *ror* stands for the weighted average of debt and equity costs, *p* is the price of the asset and *q* is the number of assets installed. Note that the impact of taxes is ignored. Furthermore, the assumption is that the complete investment is undertaken at a single moment in time. In reality, the construction time of a new network may span a significant period of time with some assets being installed earlier than others. Given however that the time scope of the analysis is very long (the lifetime of assets in electricity distribution is typically around 30 years), the impact of this assumption will be limited.

In addition to investment costs, the firm will also face the costs of network losses. Due to the inherent resistance of connections, there will be ohmic losses in the network. The losses for each connection *c* can be approximated by Willis (1997, p. 32):

$$losses_c = \hat{I}_c^{\,2} \cdot \Omega_c \cdot 8760 \cdot LF_c^{\,2} \tag{6-10}$$

Where $\hat{I}c$ stands for the peak current through the connection in normal operating conditions, Ω is the resistance of the connection, and *LF* is the load factor. The cost of losses can be calculated by multiplying the total amount of losses (in kWh) by the purchase price per kWh for these losses. Assuming a fixed price per kWh, the expression of total network cost then becomes:

$$nc = \sum_a capex_a + p_L \cdot \sum_c losses_c \tag{6-11}$$

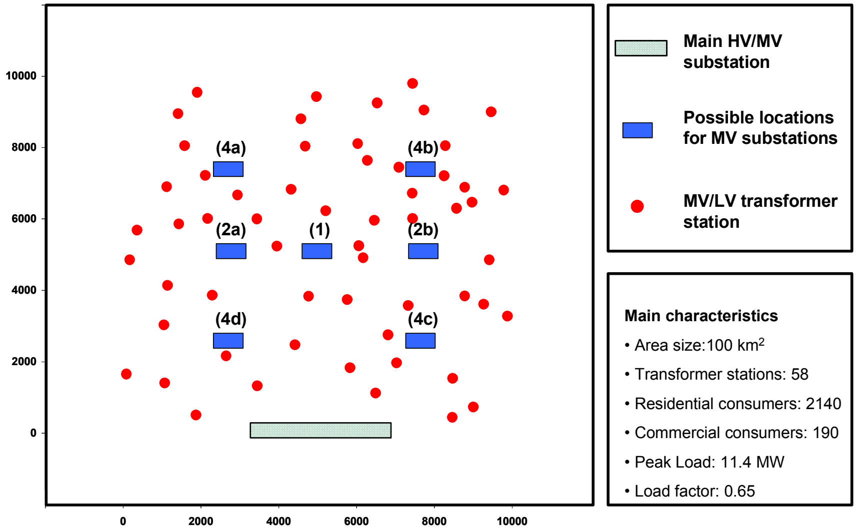
Where p_L is the price paid for one kWh of losses. The annual social cost of the network can then be defined in the following familiar fashion:

$$sotex = ic + nc \tag{6-12}$$

6.4 Case Study

6.4.1 Input Data

The NST has been applied to a fictive Greenfield supply area of 10x10 km consisting of 2330 consumers (represented by 58 MV/LV transformer stations) serving clusters of residential and commercial consumers. Figure 6-5 provides a schematic overview of the Greenfield supply area.



Main characteristics

- Area size:100 km^2
- Transformer stations: 58
- Residential consumers: 2140
- Commercial consumers: 190
- Peak Load: 11.4 MW
- Load factor: 0.65

Figure 6-5. Overview of Greenfield supply area. Distances are in metres

As the analysis deals with distribution networks only, industrial consumers are excluded from the analysis. Residential consumers and commercial consumers are supplied by 250 kVA and 400 kVA transformers, respectively. A peak load of 5 kVA and 40 kVA for these two types of consumers is assumed. In total, the distribution network serves 2140 residential and 190 commercial consumers. This results in a non-coincidental peak load of around 15 MVA. In reality, not all consumers will demand their peak load at the same time. To capture this effect, the non-coincident peak load is multiplied by a coincidence factor of 0.8, which brings the coincident peak load to approximately 12 MVA. Furthermore, a power factor of 0.95 is assumed, which results in an active power coincidental peak of 11.4 MW. Throughout the analysis, a load factor of 0.65 is assumed.

The MV/LV transformers are supplied via the MV substations and are defined as input data for the model. The number of MV substations is a decision variable in the analysis and can be either one, two or four. The location of these MV substations depends on this number; the different options are shown in Figure 6-5 for each of the three respective cases of one, two or four MV transformers. Each MV substation in turn is supplied by a single main HV/MV substation the location of which is fixed at 5000 metres at the South of the centre of the supply area. The network consists of underground cables only. The assumption is that each two kilometres of cable contains one joint. Joints connect two parts of a cable within the network and are also subject to failure. Thus, the inclusion of joints decreases the reliability of the network.

Information on prices for network components has been obtained from Ajodhia (1999) and inflated to 2005 levels. A rate-of-return of 10 percent and a lifetime of 30 years has been assumed for all assets. Table 6-4 shows the price of the different network components as well as the value assigned to network losses. Component failure data are obtained from Dutch statistics and are based on a compilation by Meeuwsen (1998b).

Table 6-4. All-in prices for different network components and network losses joints used in the case study.

Component	Price	Unit
MV substation building	70,000	EUR / station
Network Cable-AL95	45	EUR / m
Network Cable-AL240	80	EUR / m
Network Cable-AL400	90	EUR / m
Network Cable-AL630	130	EUR / m
Primary feeder breaker	39,000	EUR / feeder
Secondary feeders breaker	11,000	EUR / feeder
Feeder protection	8,000	EUR / feeder
Disconnector	4,000	EUR / disconnector
Network Losses	35	EUR / MWh

Table 6-5 shows the failure rates and repair and switching times assumed for cables and joints. All other components (e.g. breakers, rail sections, and protection systems) are assumed perfectly reliable.

Table 6-5. Failure data for cables and joints used in the case study.

Component	Failure Rate	Repair Time	Switching time
Cables	0.0018 outages / km per year	24 hours	1 hour
Joints	0.0023 outages per year	8 hours	N/A

For the calculation of interruption costs, use has been made of the value of non-delivered energy as applied by the Norwegian regulator NVE (Langset et al. 2001). A distinction is made between residential and commercial consumers. Table 6-6 shows the cost per kWh of non-supplied electricity for each type of consumer.

Table 6-6. Interruption cost data used in the case study.

Consumer Type	Interruption costs per kWh non-supplied
Residential	0.40 EUR / kWh
Commercial	4.50 EUR / kWh

6.4.2 Simulation Results

Total calculation time for the case study is about one hour when the model runs on a PC with a Pentium 4 (1.8 GHz) processor. The result of the analysis for the given Greenfield supply area is shown in Figure 6-6, which displays the annual sotex for each network alternative as a function of quality (measured by SAIDI).

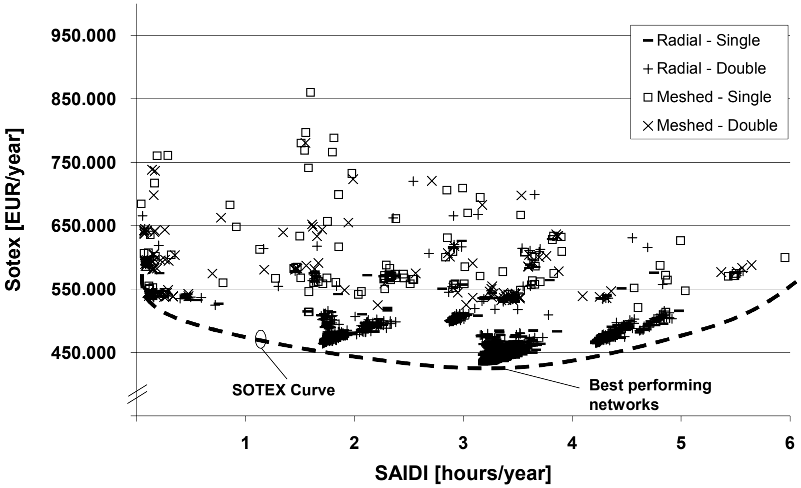


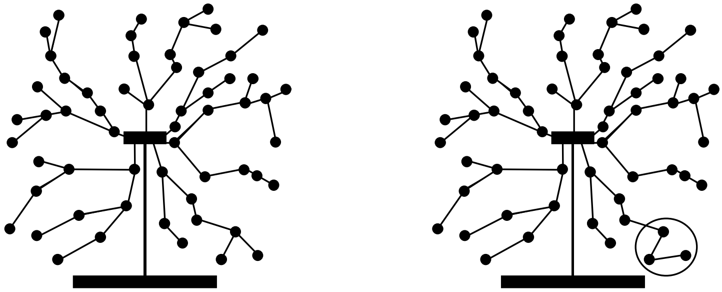
Figure 6-6. Sotex and quality performance of each (type of) network alternative.

Figure 6-6 also provides an approximation of the sotex curve, which can be drawn by enveloping the points representing those networks which exhibit the lowest sotex for a given quality level. The form of the curve is in line with expectations. Starting from a high quality level (SAIDI=0) and moving to the right (higher SAIDI levels), sotex levels first decrease and then increase once the optimal quality level is reached. The sotex curve as drawn here should, however, be interpreted with caution since its shape is biased by the genetic algorithm. In reality, the sotex curve is likely to be less steep than shown in Figure 6-6.

As can be observed, the best performing network alternative has a SAIDI performance of about 3.2 hours per year and an annual sotex of around 430,000 EUR per year. Figure 6-7(a) shows a schematic representation of the best performing network alternative. As may be observed, this network has a radial design and consists of a single MV substation and a single primary feeder. In terms of social costs, however, the difference between the best performing and next best performing networks is very small. As is shown in Figure 6-7, the four best performing networks have the same main characteristics; there are only marginal differences in the routing of these networks. Inaccuracies resulting from uncertainty about the input data and modelling assumptions are likely to eclipse such small differences. Also, due to the partially stochastic nature of the search process, the optimal network, i.e. the best performing network alternative will in principle always be (slightly) different each time the same analysis is performed. Considering these factors, it does not seem appropriate to consider the resulting best performing network as the truly optimal one. Rather, the main outcome of the analysis should be seen as an identification of the characteristics of an optimal network for the case at hand.

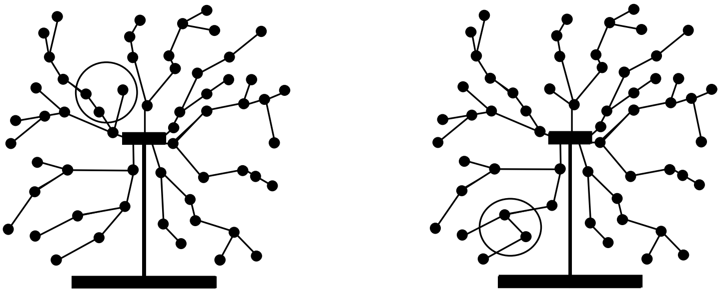
In this respect, it is helpful to make a classification of networks along two dimensions, namely whether the network consists of radial or meshed feeders, and whether a single or double primary feeder is being used. A graphical presentation of the performance of the four resulting classes of networks is provided in Figure 6-6. The average performance of each class of networks is presented in Figure 6-8. As can be seen, the quality performance of radial networks is – on average – lower than that of meshed networks. This is caused by the longer average restoration times. When a network is meshed, there is the possibility to reroute power flows if an interruption occurs. This causes the duration of an interruption to decrease. Meshed networks, however, are more expensive than radial ones. There are two main reasons for this. Firstly, a meshed design consists of more connections than a radial network. Secondly, the capacity of connections in meshed networks is typically higher than in a radial network in order to enable the rerouting of power flows in the case of a fault. If a certain feeder is rerouted, the loading of the feeder segments that remain in operation generally increases. This leads to the need to install feeders of higher capacity in meshed networks and therefore leads to higher costs.

For the given Greenfield supply area, the difference between radial and meshed networks in terms of network costs is, on average, around 150,000 EUR/year. On the other hand, the decrease in interruption costs when choosing a meshed network instead of a radial one averages only 26,000 EUR/year. Thus, for this particular supply area, it seems uneconomic to opt for a meshed design since the reduction in interruption costs does not outweigh the additional network costs. Although there may be individual meshed networks that perform better than certain radial ones, radial networks on average provide a better price-quality trade-off. Furthermore, as can be seen in Figure 6-6, networks that are close to the bottom of the sotex curve are primarily of a radial design.



(a) Best performing alternative

(b) Second best performing alternative

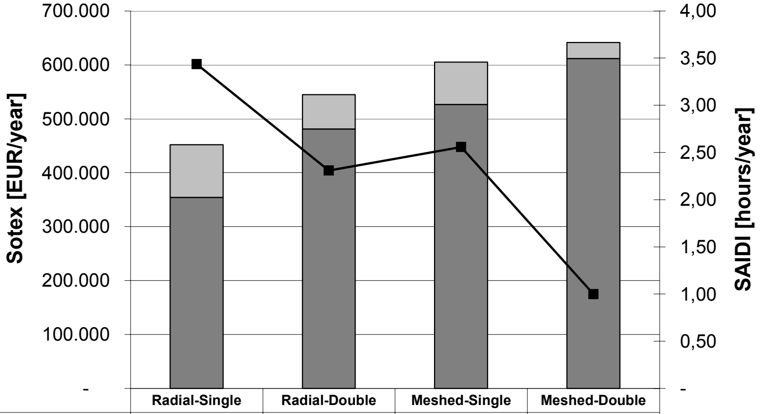


(c) Third best performing alternative

(d) Fourth best performing alternative

Alternative:	1	2	3	4
Number of MV substations	1	1	1	1
Basic Topology	Radial	Radial	Radial	Radial
Number of primary feeders per MV substation	1	1	1	1
Total number of secondary feeders	6	6	6	6
Protection Primary Feeder 1	1	1	1	1
Protection Primary Feeder 2	1	1	1	1
Protection for secondary feeders	1	1	1	1
SAIDI	3.19	3.19	3.19	3.19
Network Costs	342,108	342,009	340,091	342,285
Interruption Costs	89,217	89,395	91,314	89,134
Sotex	431,320	431,404	431,404	431,419

Figure 6-7. Overview and main characteristics of the four best-performing network. The circles indicate the difference between the best and subsequent performing networks.



	Radial-Single	Radial-Double	Meshed-Single	Meshed-Double
Sotex	452.061	544.789	605.023	641.519
□ Interruption Costs	97.627	63.677	78.441	29.761
■ Network Costs	354.434	481.112	526.581	611.759
—■— SAIDI	3,44	2,31	2,56	1,00

Figure 6-8. Summary of results for four classes of networks.

A similar comparison as above can be made between networks with a single or a double primary feeder. Installing an additional primary feeder between the main HV/MV substation and the MV substations leads to an increase in quality as faults in one of the primary feeders no longer result in an interruption. If one primary feeder fails, electricity can still be supplied through the second feeder. However, the associated decrease in interruption costs (approximately 40,000 EUR/year) is not sufficient to cover the cost of the second primary feeder (approximately 105,000 EUR/year). This suggests that, in this particular case, installing a single primary feeder is the most efficient option as this represents a better trade-off between cost and quality.

In summary, a radial network comprising a single primary feeder is the network design that best suits the given Greenfield supply area. This type of network provides the best trade-off between quality and cost. As can be seen in Figure 6-7, such a configuration results in a SAIDI value of 3.2 hours per year and a sotex value around 430,000 EUR/year. Using this information, the regulator would be able to evaluate the firm's investment proposal in a more effective way. For example, if the firm proposed a meshed network with double primary feeders, the regulator could argue that – even though that design would lead to a higher quality level – the firm's proposal will be too expensive from the social point of view. Moreover, if the main characteristics of the network proposed by the firm are similar to those of the best networks identified by the NST tool but the cost of the former is significantly higher than that of the latter, the regulator could request the firm to explain this difference.

6.4.3 Sensitivity Analysis

The outcome of the NST is conditioned by a number of assumptions regarding the input data. More specifically, prices and failure parameters of network components play a key role. The same also holds for interruption cost data. A change in these input parameters can influence the outcome of the analysis. In order to assess the sensitivity of the optimal price-quality trade-off with respect to these parameters, a number of scenarios have been analysed. Besides a base case scenario, eight additional scenarios have been created where the input parameters have been either increased or decreased by a factor of two. The input parameters, which were subject to a sensitivity analysis, include the prices of network components, the failure rate of cables and joints, the repair and switching time of cables and joints, and the interruption costs per kWh. Figures 6-9 and 6-10 graphically illustrate the impact of the change in input parameters on the price-quality trade-off. An overview of key results, as compared to those for the base scenario, is shown in Table 6-7. A discussion of the results now follows.

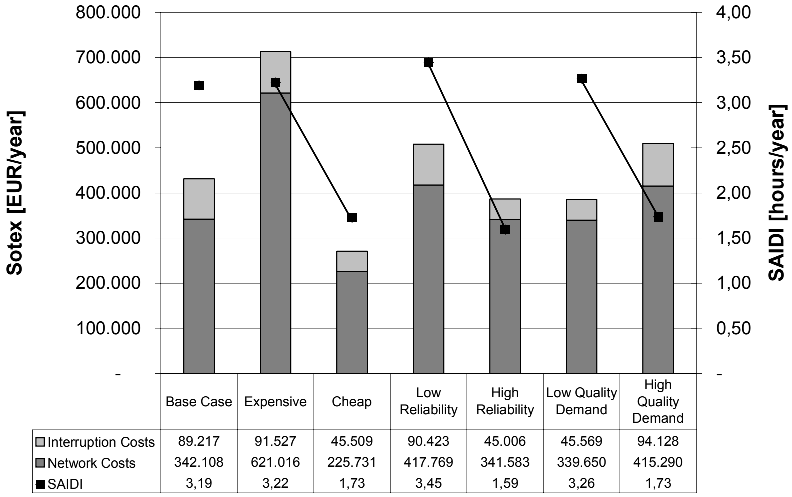
Decreasing the price of network equipment leads to an increase in the optimal quality level i.e. a lower SAIDI level. If prices are low enough, it becomes economical to install additional primary feeders. This leads to an increase in reliability since interruptions affecting a primary feeder no longer result in interruptions. A decrease in equipment prices thus leads to an increase in the optimal quality level. From a more general point of view, reducing network costs leads to socially better outcomes, which comes in the form of lower prices and higher quality levels. If the firm is able to operate in a more efficient manner, the cost of adding redundancy to the network decreases and consequently quality goes up. In the opposite case, i.e. if the firm is operating less efficiently, then it becomes more expensive to provide a higher quality level. Overall, consumers then pay more while they are provided with lower levels of quality.

In this particular case, an increase in equipment prices has little impact on the optimal quality level relative to the base case. The explanation for this is the limited possibility to further reduce the redundancy of the network. The main cost saving measure that could be taken is removing protections from feeders. This, however, has a substantial negative impact on quality whereas the associated cost savings are relatively low. As may be observed, doubling component prices does not justify savings on protection as the resulting increase in interruption costs would then exceed the decrease in network construction, operation and maintenance costs.

Increasing the failure rate and repair and switching times of network components leads to respectively more frequent and longer interruptions. At the one hand, a higher failure rate implies that components fail more frequently and therefore, the probability of an interruption to occur also increases. Increasing the repair and switching time of components means that it now takes longer to restore supply either by repairing the component or through rerouting measures. Less reliable components thus lead to an increase in interruption costs. Being confronted with low quality components, at some point it becomes economic to increase the redundancy by installing an additional primary feeder.

Scenario	Base Case	Equipment Price		Failure Rate		Repair / Switching Time		Interruption Costs	
		Expensive	Cheap	Low Reliability	High Reliability	Low Reliability	High Reliability	Low Quality Demand	High Quality Demand
Nr. of MV substations	1	1	1	1	1	1	1	1	1
Basic Topology	Radial	Radial	Radial	Radial	Radial	Radial	Radial	Radial	Radial
Nr. of primary feeders per MV substation	1	1	2	2	1	2	1	1	2
Nr. of secondary feeders	6	6	7	7	6	7	6	5	7
Protection Primary Feeder 1	1	1	1	1	1	1	1	1	1
Protection Primary Feeder 2	1	1	1	1	1	1	1	1	1
Protection for secondary feeders	1	1	1	1	1	1	1	1	1
SAIDI [hours/year]	3.19	3.22	1.73	3.42	1.60	3.47	1.59	3.26	1.73
Network Costs [EUR/year]	342,108	621,016	225,731	416,985	341,078	418,552	342,087	339,650	415,290
Interruption Costs [EUR/year]	89,217	91,527	45,509	90,208	45,394	90,638	44,617	45,569	94,128
Sotex [EUR/year]	431,320	712,550	271,240	507,190	386,480	509,190	386,700	385,220	509,420

Table 6-7. Performance of best performing networks under different scenarios.



	Base Case	Expensive	Cheap	Low Reliability	High Reliability	Low Quality Demand	High Quality Demand
☐ Interruption Costs	89.217	91.527	45.509	90.423	45.006	45.569	94.128
▨ Network Costs	342.108	621.016	225.731	417.769	341.583	339.650	415.290
■ SAIDI	3,19	3,22	1,73	3,45	1,59	3,26	1,73

Figure 6-9. Optimal price-quality trade-off under different scenarios.

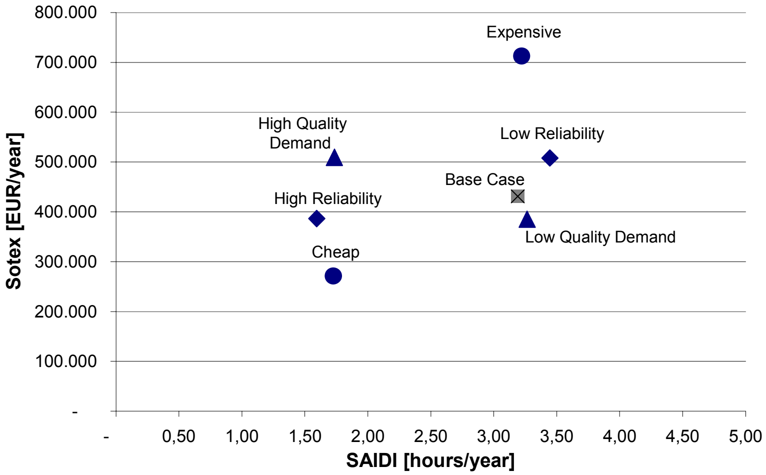


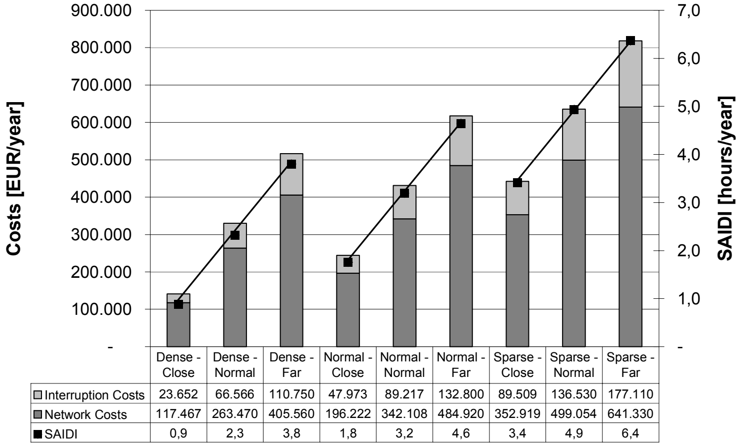
Figure 6-10. Graphical representation of the optimal price-quality trade-offs under different scenarios.

This can be interpreted as follows. The primary feeder has substantial impact on the overall quality of the network as a fault in this feeder results in an interruption affecting all consumers. If the reliability of the primary feeder decreases, then the frequency and duration of faults affecting this feeder increases accordingly. Consequently, interruption costs increase and it becomes more economic to invest in a second primary feeder. This leads to a substantial improvement in service quality as a fault in a single feeder no longer causes interruptions. In this particular case, net savings are high enough to justify the investment in an additional primary feeder.

Overall, however, consumers are still worse off if component reliability is lower. On the one side, the optimal quality level is reduced (due to less reliable network components) while on the other side, network costs increase due to the investment in an additional primary feeder. This is in line with expectations since the use of less reliable network components (at constant prices) makes the provision of quality more expensive; this implies a lower optimal quality level. Conversely, the use of more reliable network components – represented by a decrease in the failure rate and repair and switching times – leads to an increase in the optimal quality level while total network costs remain more or less the same. At the same time, interruption costs are now lower and this leads to a net reduction in the total level of sotex. Consumers thus benefit in the sense that they are provided with a higher quality service for the same cost. Note also that the use of high quality components reduces the need to add additional redundancy to the network. For example, doubling the number of primary feeders leads to a relatively small increase in quality if the reliability of the feeder is very high. The additional feeder would however still come at significant costs.

A change in consumer demand for quality is reflected in the level of interruption costs per kWh of non-supplied energy. An increase in quality demand can thus be approximated by increasing the level of interruption costs per kWh. As consumers place a higher value on quality, it becomes more economical to improve quality levels as the costs to do so are less than the resulting benefits. In this case, the cost of adding a second primary feeder leads to a reduction in interruption costs, which justifies the cost of that extra feeder. The increasing demand for quality justifies the fact that now consumers have to pay a higher price for the distribution service. On the other hand, if interruption costs decrease, the optimal quality level and network costs decrease as well thus leading to a net improvement in sotex levels.

Apart from assessing the impact of changing parameter values on the outcome of the analysis, it is also interesting to look at the impact of changes in the structural characteristics of the supply area itself. Two factors have been considered, namely the density of demand and the location of the main HV/MV substation. With respect to demand density, two cases can be identified being a “dense” network and a “sparse” network. These two cases correspond to a situation where the size of the supply area has been respectively contracted and expanded by a factor two. With respect to the location of the HV/MV substation, the distance of this substation to the centre of the supply area has been modified. In the base case, the HV/MV substation is located 5,000 metres South of the centre of the supply area. For the purpose of the sensitivity analysis, this distance has been changed to 10,000 metres in one case and zero metres in the other. These two cases are denoted as “close” and “far”, respectively. Only the vertical position of the substation has been modified; the horizontal position has been maintained at the centre of the supply area.



	Dense - Close	Dense - Normal	Dense - Far	Normal - Close	Normal - Normal	Normal - Far	Sparse - Close	Sparse - Normal	Sparse - Far
□ Interruption Costs	23.652	66.566	110.750	47.973	89.217	132.800	89.509	136.530	177.110
▨ Network Costs	117.467	263.470	405.560	196.222	342.108	484.920	352.919	499.054	641.330
■ SAIDI	0,9	2,3	3,8	1,8	3,2	4,6	3,4	4,9	6,4

Figure 6-11. Results for the best performing network alternative under different density and distance scenarios.

The results reveal that changing the load pattern or the substation position does not alter the basic characteristics of the optimal network for the Greenfield area under consideration. The best performing alternative remains a radial network with a single primary feeder in all cases. However, the changes do have some impact on the trade-off between cost and quality. As shown in Figure 6-11, moving the HV/MV substation further away from the supply area centre leads to higher network costs as well as higher interruption costs. Longer distances imply longer cables and therefore higher cable costs. At the same time, the probability of a failure increases with the length of the cables and therefore leads to a decrease in service quality. It remains uneconomical, however, to install an extra primary feeder. The costs of this extra feeder are still higher than the accrued benefits in terms of a decrease in interruption costs.

The effect of demand density on results has also been assessed. Demand density can be defined as the average distance between loads and the centre of the supply area. As can be observed in Figure 6-11, network costs decrease in the case of a dense network as the average length of feeders is reduced. At the same time, shorter feeders also lead to an improvement in quality. Sparse networks on the other hand are more expensive due to the longer distances that feeders need to cover. In addition, there is a decrease in quality as the probability of a feeder being exposed to a fault is also higher due to longer feeder length. A supply area with high load density is thus more attractive to serve as this can be accomplished at relatively low costs while at the same time, a higher quality level can be obtained.

6.5 Conclusions

This chapter has developed a new methodology for evaluating the combined price and quality performance of a new network investment under the building blocks framework. The basic idea is to identify the best possible network design for a given supply area and use this as the benchmark for evaluating the distribution firm's own investment proposal. The results of the case study indicate that the NST may well be an important tool in the process of investment appraisal. The NST can provide the regulator with valuable information regarding the preferred network characteristics for the given supply area. What is more, it allows the regulator to analyse the impact of changes in input parameters and supply area characteristics on the price-quality trade-off. In doing so, the NST takes into account possible spatial differences in the demand for quality across the network. The resulting optimal network reflects a trade-off between price and quality for the system as a whole but, at the same time, considers the fact that consumers located in different places will place different values on the quality of service. Lastly, instead of considering cost and quality separately as other models do, the NST adopts an integrated approach in making the trade-off between these two aspects of network performance.

The limitations of the NST should however also be recognised. The NST is only a model and therefore an imperfect representation of reality. The outcome of the analysis is driven by the assumptions made in the modelling process. Similarly as with optimisation tools, simulation approaches like the NST run the risk of arriving at a local rather than a global optimum. Making fewer assumptions increases the probability of identifying the true optimal network but at the same time, also leads to longer computation times. Thus, achieving a trade-off between the realism and the practicality of the analysis is essential. In the face of the modelling restrictions, it is unlikely that the best network identified by the NST will coincide with the true optimal network. Furthermore, uncertainties in parameters as well as in the input data may adversely influence the NST outcome.

Recognising these limitations, it is more appropriate to consider the NST as a tool aimed at revealing information about a range of investments and the desirable properties of a network rather than providing an exact prescription of what should be the optimal network for a given supply area. For example, the NST results should not be used to discuss specific routing details of the network. Such specific considerations would also not fit within a price-cap framework and tend to lead to undesirable micro-management by the regulator. Rather, the detailed technical design of the network should be left at the firm's discretion as long as it complies with the general constraints imposed on the basis of the information revealed by the NST.

Rather than getting involved in the details of the firm's operations, price-cap regulation aims to provide socially desirable outcomes by means of sending incentives. In the spirit of price-cap regulation, it is not desirable for the regulator to intervene in the investment planning process of the firm but rather, to set price and quality targets that are considered appropriate from a social point of view. The NST provides regulators with an – albeit imperfect – instrument to obtain information about the performance of a given investment proposal. This can be realised without the need to perform detailed engineering assessments while the amount of data required for the analysis is relatively modest. The information revealed by the NST can then be fed into the process of assessing the desirability of investments or, more generally, into the determination of the firm's Regulated Asset Base.

Notes

[1] At the time of writing, the Swedish regulator had not yet formally implemented the NPAM. See Gammelgard and Larsson (2003) or Larsson (2003) for a description of the Swedish network model.

[2] For an overview of optimisation algorithms in distribution network planning see for example Khator and Leung (1997) or Brown (2002) pp. 291-353.

[3] In genetic terms, this string of bits could be considered the network's chromosome.

[4] See also Brown (2002, p. 311) for a description of these parental selection methods.

[5] For a detailed treatment of electricity network reliability analysis, see for example Meeuwsen (1998a).

[6] Distrel was developed by the Power Systems Laboratory of the Delft University of Technology. See Meeuwsen (1998b) for a detailed description of Distrel.

[7] In the case of transmission networks, the analysis of multi-order faults would be more important as the level of redundancy in these networks is typically higher.

7

Conclusions and Recommendations

7.1 Introduction

There is a trend towards stricter forms of price control of electricity distribution networks; this new approach is generally known as price-cap regulation. Price-caps are intended to ensure higher efficiency and ultimately, lower prices for consumers. However, theory as well as empirical evidence reveal a risk of adverse quality reduction under price-caps. Undesired reductions in the level of quality may well overshadow the efficiency benefits generated by the price-cap. At the same time, society is becoming increasingly dependent on a reliable supply of electricity. Interruptions in supply have significant economic impact. The promise of price-cap regulation can hence only be fulfilled if explicit provisions for quality are also set in place. This entails a need to establish an integrated approach towards price and quality regulation – a challenge that has been taken up in this thesis.

This thesis was focused on the development of an integrated approach for optimal price-quality regulation of electricity distribution networks under a system of price-caps. An important feature of this thesis is that it considers price and quality from within an integrated framework. Traditionally, regulators have tended to approach the quality issue under price-caps from an isolated point of view. This is reflected in the determination of the efficiency improvement potential factor (the 'X-factor'), which tends to focus on costs alone, whilst quality is usually regulated on top through one or more quality controls. In contrast, this thesis focuses on the issue of how to determine the integrated price-cap as featured by a truly integrated X-factor – one that incorporates improvement potential in the realms of both price (efficiency) and quality.

This thesis explored the quality problem under price-cap regulation and developed two approaches for integrated price-quality regulation and benchmarking. Chapter two started by reviewing the literature on price and quality regulation of monopoly. At the centre of attention was the question of the optimal price and quality level. Chapter three then continued by developing a taxonomy of price-cap approaches. This resulted in the identification of two main regulatory approaches namely the totex and the building blocks approach. In the context of an integrated price-quality approach, the optimal X-factor does not only reflect scope for improvement towards an efficient level of costs, but also towards the optimal level of quality. The problems related to identifying this optimal quality level were assessed in chapter four. The ability to set the optimal X-factor, taking into account both price and quality aspects, largely depends on the ability of the regulator to successfully overcome his inferior informational position. As part of this process, benchmarking is an important regulatory asset. Chapters five and six developed benchmarking models for the two regulatory approaches. These novel benchmarking methods are based on the idea of assessing price and quality performance in an integrated way. Taking into account their limitations, these models have shown to be able to extract valuable information that can subsequently be used by the regulator to improve the effectiveness of his regulatory approach.

This chapter presents the conclusions and recommendations for further research. A discussion of the contributions of this thesis now follows along the lines of the three research questions as formulated in the first chapter. Then, some policy implications are discussed and recommendations for further research are provided.

7.2 Research Synthesis

7.2.1 Research Question 1: Optimal Price and Quality

The first research question deals with the issue of the optimal price and quality level. When moving to price-cap regulation, both theory and empirical evidence show that there is an explicit regulatory need to consider quality under the price-cap. Fundamental for successfully doing so however, is a better understanding of what regulators should aim at, i.e. what defines the optimal price and quality pair. This thesis reviewed the theory underlying price and quality regulation under a monopoly regime and identified the problems experienced in attaining the optimal price and quality outcome.

An optimal price-quality pair complies with two basic requirements. Firstly, regarding the price aspect, it implies that the firm is producing in an efficient manner. That is, the given outputs are produced using the least possible level and best possible combination of inputs. Secondly, regarding the quality aspect, it means that the firm provides a level of quality that is in line with consumer demand. In practical terms, this means that the level of total social costs – as measured by the sum of network costs and interruption costs – is minimised. The objective of an integrated price-cap system can then be defined as achieving this socially optimal price and quality pair.

In pursuing this objective, the regulator is hindered by lack of information. In the context of integrated price-quality regulation, this informational problem may be decomposed into three issues: (1) the ability to measure quality, (2) the ability to measure consumer demand for quality, and (3) knowledge of the cost-quality relationship. Dealing effectively with these three problems is the key to regulatory success.

Quality measurement, which is the first issue, requires the prescription and implementation of adequate quality monitoring and registration systems. The second problem, measuring consumer demand for quality, takes the form of measuring consumer interruption costs. These interruption costs can differ as a function of many variables such as the type of consumer, the timing of the interruption, and the duration of the interruption. The large variety in interruption characteristics forces one to adopt some average measure for interruption costs, possibly differentiated by consumer class. For dealing with the third informational problem, lack of information about the cost-quality relationship, benchmarking is an important regulatory asset. By comparing the relative price and quality performance of firms, the regulator can develop a better understanding of the improvement potential of these firms and consequently impose adequate improvement targets. Incorporation of integrated price-quality benchmarking models into the determination of the X-factor can thus lead to a more effective price-cap system.

7.2.2 Research Question 2: Integrated Price-Quality Regulation

The second research question asks what approaches can be followed to design price and quality regulation and to what extent these approaches are effective. This thesis has developed a taxonomy of price-cap approaches and evaluated these with respect to their price and quality performance. A distinction can be made between four main regulatory strategies namely isolated caps, related caps, yardstick competition, and sliding scales. These strategies differ in two respects: The time of setting the X-factor, and the extent to which the X-factors of firms are related to each other. Depending on how these strategies are applied to the firm's costs (opex and capex), two main approaches can be identified.

The first approach is known as the totex approach. Here, the regulator applies the same price-cap strategy to the firm's total costs i.e. the sum of operational and capital expenditures (opex and capex). The regulator makes no distinction between opex and capex and applies the X-factor on the basis of realised total cost levels. If a given firm manages to reduce total costs (more than others do), it earns higher profits both in the given regulatory period and in the next period, as prices are based on cost performance in the previous period.

The second approach is the building blocks approach. Here, the regulator makes a distinction between opex and capex and regulates these costs, in principle, based on different price-cap strategies. Allowances for opex are typically set on the basis of an opex-only benchmarking analysis. Regulation of capex on the other hand takes the form of assessing the firm's investment proposals in order to determine whether allowances will be given for depreciation and return of these investments.

The two approaches differ with respect to their price (efficiency) and quality effectiveness. Regulatory effectiveness, in this context, can be defined as the extent to which the chosen approach leads to higher productivity as well as an optimal quality level. The totex approach provides strong efficiency incentives but entails a higher risk of quality degradation. Conversely, the building blocks approach provides weaker efficiency incentives but provides better possibilities to control the quality outcome.

The totex approach leads to strong efficiency incentives as it links future prices to past total cost performance. Furthermore, it does not allow for suboptimal allocations between opex and capex. However, the quality problem is also more relevant as here, the firm is in principle free to choose its spending levels – including its investment level. This may be particularly problematic given that quality in electricity distribution is strongly capital driven. Postponing investments or even refraining from investing has a positive effect on profitability in the short term but is likely to lead to adverse effects on quality in the longer term. The building blocks approach on the other hand is primarily focused on incentives in the area of opex. Efficiency incentives for capex are limited and allowances for investments are effectively set on the basis of the firm's own projections. However, the building blocks approach is also associated with a lower risk of quality degradation as investment levels can be indirectly controlled by the regulator.

In line with the two price-cap approaches, two approaches may be identified for integrating quality into the price-cap. Under the totex approach, quality integration takes the form of an integrated assessment of previous performance of the firm. Rather than only basing the X-factor on past cost performance, the X-factor is now set on the basis of combined price and quality performance as featured by the total social costs resulting from the firm's cost and quality decisions. Firms that manage to make a better price-quality trade-off will incur less sotex and therefore gain a higher efficiency score. Under the building blocks approach, integrated regulation takes the form of a combined price-quality assessment of proposed investments. Here, the regulator should make sure that investments that are allowed into the firm's capital base – and that will therefore ultimately be reflected in the allowed price – are those associated with least levels of sotex i.e. are implemented at a cost level that reflects an efficient mode of production and that provides a level of quality that is optimal.

7.2.3 Research Question 3: Integrated Price-Quality Benchmarking

The third and final research question was concerned with the measures that could be designed or developed to improve the effectiveness of the two regulatory approaches. Under both the totex and building blocks approach, the regulatory ability to successfully evaluate price-quality performance is crucial to the effectiveness of the approach. The better information the regulator has available, the better the appropriate X-factor can be estimated and the more effective will be the regulatory approach. On the one hand, consumers will then directly share in the anticipated improvements while, on the other hand, the regulator can make sure that the improvement targets imposed on the firm are realistic and do not lead to financial sustainability problems.

If the regulator is better informed, he can establish more effective incentives i.e. assure that the firm operates in a socially optimal way as well as assure that consumers share in the benefits resulting from the increase in social welfare. Benchmarking is an important regulatory asset in acquiring information about the firm's performance. Traditionally, benchmarking under price-caps has been mostly restricted to cost efficiency assessment. A truly integrated benchmarking that considers both costs and quality was lacking.

An integrated X-factor reflects the regulatory anticipation of both price and quality improvements. The ability to measure productivity and quality performance in terms of a single indicator is therefore an important regulatory asset. Integrated benchmarking deals with the problem of measuring the cost-quality relationship. In practical terms, combined price and quality performance takes the form of measuring total social costs or sotex, being the sum of network costs and consumer interruption costs. Two benchmarking methodologies for measuring sotex performance under the two respective price-cap approaches have been developed in this thesis. As demonstrated through empirical verification, application of the two benchmarking methods can help the regulator to acquire crucial information about the desired price and quality levels. This information can subsequently be used to design more effective price-cap systems.

At the same time, however, the limitations of benchmarking should also be recognised. These limitations stem from the three informational problems. If the regulator cannot properly measure costs, quality, or consumer demand for quality, the results of the benchmarking analysis can be adversely affected. Furthermore, the benchmarking outcome is driven by modelling assumptions regarding the cost-quality relationship – these assumptions are bound to be imperfect. For benchmarking to be effective, these limitations should be recognised and properly dealt with in the determination of the integrated X-factor. The role of benchmarking tools therefore is primarily one of assisting the regulator in extracting (albeit imperfect) information from the firm in order to allow him to estimate the range within which the X-factor should fall. The better the regulator is able to conduct the benchmarking analysis, the narrower this range will be and the more accurately the X-factor can be set. Benchmarking thus is not the ultimate solution to the regulator's informational disadvantage, but rather a tool to support the regulator in dealing intelligently with the informational problem.

7.3 Policy Implications

7.3.1 Privatisation of Distribution Networks

Apart from contributing to the body of academic knowledge, the results of this research may also be used to feed into some of the contemporary discussions in the field of power sector reform. In some countries, and in particular the Netherlands, there is an on-going debate on the necessity of privatisation of electricity distribution networks. Some advocate that these networks should be kept in public ownership as they fulfil a crucial role in the supply of electricity; society can simply not run the risk to give away this responsibility. Others feel that privatisation will generate additional benefits in terms of e.g. higher productivity.

From the quality perspective, there are two separate but related questions that the results of this research may help to address. Firstly, there is the question whether one should privatise electricity distribution networks or not. Generally, one may expect a privately owned distribution firm to operate more efficiently than a public one. However, as empirical evidence points out, this is not necessarily true in the case of a monopoly. In a major international power sector study, no concluding evidence was found that private monopolists operate more efficiently than public ones (Pollit 1995). Higher productivity seems to be driven by the presence of competition rather than the type of ownership.

This observation presents a strong counter-argument for privatising distribution network monopolies – in particular when the quality dimension is taken into account. Not only are the associated efficiency benefits disputable, the risks of adverse quality effects under private ownership are likely to be higher. Private network owners may be assumed to have a stronger profit (but not necessarily efficiency) incentive than their public counterparts, and might hence be more inclined to reduce costs at the expense of quality. Public ownership, in the presence of the appropriate democratic institutions, can act as a form of insurance against a destructive quality policy. Public ownership itself may indeed be considered as a form of regulation i.e. a means to safeguard public interests. This insurance is lacking under private ownership, which implies a higher probability of exploitation of monopoly power and therefore, among others, an increased risk of adverse quality effects. The quality problem is likely to be exacerbated in the case of a private monopolist. This implies a need for a stronger regulator, both in the quantitative and qualitative sense. Clearly, the ensuing additional regulatory costs should be taken into account when evaluating the costs and benefits of privatisation.

The second question relates to the form of the privatisation. In the Netherlands, for example, there is a debate whether distribution networks should be privatised in isolation or in combination with the associated (incumbent) supply businesses. Network and supply business are already unbundled at the accounting and management level; not yet at the ownership level. As highlighted in section 4.2.1, and empirically verified in the Italian case study (Annex IV), the monopoly distributor has an indirect incentive to supply a high level of quality to its consumers if these consumers are also clients of associated supply or generation firms operating in competitive markets. This spill-over effect is particularly strong if these associated firms operate under the same brand name. This effect acts as an incentive to the distribution firm to take into account consumer demands in the determination of the quality level. At the same time, this implies a reduced need for explicit regulatory quality intervention and thus lower regulatory costs.

Clearly, the spill-over effect would only be present in the presence of competitive pressure, i.e. if liberalisation has actually resulted into effective competition. Thus, although the spill-over effect may on the one side reduce the need for regulation of the network monopoly business, it may, on the other side, provide the incumbent supplier an unfair advantage over other suppliers also wishing to enter the market. From this perspective, joint ownership of the monopoly network and the competitive supply (as well as other parts of the) business would therefore be undesirable. At the same time, separate ownership implies that the spill-over effect cannot be utilised and additional quality regulation is needed. This creates an interesting dilemma.

7.3.2 Applicability of Benchmarking Tools

This thesis has developed two tools for conducting integrated price-quality benchmarking analysis. Results from the research do suggest that these tools may play an important role if applied by regulators in their daily practice. In order to move from the academic to the applied sphere, however, at least two aspects would need to be taken into account.

Firstly, serious efforts should be put in the area of data collection. If data quality is not assured, then the accuracy of benchmarking outcomes (the efficiency scores) is at least questionable. Failure to recognise the importance of data limitations in benchmarking can lead to problematic outcomes (see for example the Dutch case in Box 3-1). In particular when capital expenditures are included in the benchmark, assuring uniformity across the data sample is of the utmost importance. Furthermore, collection of quality performance data suffers from differences in definitions and measurement practices across firms as well as from different conditions between different regions served by the same firm. As long as the regulator does not have available high-quality data on costs as well as quality performance, the effectiveness of benchmarking should be seriously questioned. A first step in setting up an effective regulatory framework is therefore in the establishment of a regulatory measurement system. This would, among others, include prescriptions for cost definitions, accounting guidelines, and regulatory reporting as well as technical indicators, definitions, and guidelines for measuring quality performance.

Secondly, there is a need to recognise the multi-disciplinary nature of the regulatory problem. As was indicated in the first chapter, regulation of electricity distribution networks can benefit from a joint economics-engineering approach. Economic theory provides the framework for regulation – the use of incentives and benchmarking. Engineering insight into the technical complexities of distribution network management is needed to shape the incentives in such a way that the development of the physical network is steered in respect of public interests. Joining economic and engineering knowledge can thus help to design more effective regulatory systems. This joining of forces should not only apply in the regulatory (benchmarking) tools but also be reflected in the composition and training of regulatory staff.

In summary, the extent to which benchmarking results can actually play a role in the regulatory decision-making process will strongly depend on the credibility of the benchmarking model and underlying data as well as the presence of well-equipped regulatory staff. The work in this thesis has provided a first step in the development of integrated price-quality benchmarking tools. Evidently, however, a number of issues still need to be addressed. For example, in the specification of the benchmarking model in chapter five, no attention has been given to the presence of so-called environmental factors that may (either positively or negatively) affect the performance of a firm. Such factors include consumer density, (sub)soil composition, climate, vulnerability to natural disasters such as severe storms, floods, etc. Environmental factors are difficult to capture, in particular in the face of sample size limitations and data collection constraints. A possibility to deal with these issues is to allow distribution firms which are confronted with peculiar circumstances a one-off financial allowance. This avoids the need to increase the number of variables within the model which would decrease the discriminative power of the model. Clearly, this shifts the problem to the question of which firms should receive such one-off compensations and the level of these compensations. Detailed individual assessments or technical audits – even though this may not seem to be in line with the spirit of arm's length regulation – seem to be useful in solving this problem.

With respect to the Network Simulation Tool developed in chapter six, possible obstacles to its practical application are primarily found in the extent to which the model reflects reality. Possible routes to extend the NST are discussed in the next section.

7.4 Recommendations and Future Work

7.4.1 Further Development of the Network Simulation Tool

The scope of the NST is limited to new distribution networks, starting from a Greenfield area. In practice, firms are frequently confronted with investments that deal with the extension of existing networks or the replacement or upgrade of existing assets (e.g. for quality improvement purposes). In making the optimal investment choice, the basic principle of attaining a situation with lowest possible sotex level still holds. For example, the decision to replace a certain asset by a new version of higher quality or to increase capacity in the network is only justified if the foreseeable decrease in interruption costs exceeds the costs of doing so. The NST could be extended to also analyse the desirability of decisions such as network extensions or quality upgrade efforts from a social point of view.

Another route for extending the NST is by increasing the number of decision variables. This allows the treatment of more types of networks and improves the degree of realism of the NST. For example, in practice one can often find a network with a mixture of radial and meshed feeders. In addition, installing secondary feeders with their starting and ending point in different substations leads to higher quality as interruptions due to faults in the primary feeders can be solved relatively fast by rerouting power to the affected substation via the healthy substations. Similarly, inclusion of the impact of faults occurring in breakers or protection systems can have impact on the quality performance of the network and should also be considered in future research.

Another area of improvement concerns the scope of the NST. To start with, the scope of the NST might well be extended to transmission networks. With respect to distribution networks, the trend towards increasing numbers of dispersed generation units feeding into the distribution networks should in future be taken into account. Dispersed generation results in bi-directional rather than uni-directional flows of power in the network. This has an important impact on the cost of the network. At the same time, dispersed generation in the network may also lead to an increase in reliability as now, local generating power becomes available. Clearly, these implications will impact the optimal price-quality trade-off and would need to be considered within the NST.

Finally, the NST at this time only consider the costs of investments (including network losses) but ignores the maintenance costs. There exists a trade-off between maintenance costs and capex. For example, purchasing a more expensive but higher-quality asset reduces the maintenance costs of the asset in the longer term. This trade-off is currently lacking in the modelling of the NST.

7.4.2 Regulation of Power Quality

This thesis has only considered the reliability dimension of electricity distribution quality. In recent years, however, power quality has gained increasing attention, as the vulnerability of consumers is not only related to interruptions, but also to the physical quality of the electricity provided to them. Power quality phenomena such as voltage dips can lead to substantial consumer damage and therefore should in principle be included in the definition of social costs. The technical nature of power quality problems is different from reliability. Reliability relates to the two extremes of power quality namely electricity is available or it is not. Power quality on the other hand covers a spectrum of possibilities between these two extremes and consumers may – since the effects are not always directly visible – not be fully aware whether or not the level of power quality they are being supplied with is adequate.

In this light, future research on the topic of regulation of power quality regulation under price-caps is recommended. In order to do so, three main steps should be followed – in line with the findings of this thesis. Firstly, the research should define power quality in more specific terms, identify indicators and choose the relevant dimensions of power quality for further study. Secondly, the issue of demand for power quality should be considered. Similar to interruption cost studies, the research should analyse the different possibilities to classify and measure costs resulting from power quality problems for different consumer groups and analyse which factors influence these costs. Finally, the relation between power quality and costs should be assessed and possibilities considered for integrating power quality incentives into the price-cap. The models and approaches offered in this thesis provide the starting point that regulators may use in their difficult task to assess the appropriate X-factor under the truly integrated price-cap.

References

ACCC (2003), "Statement of Principles for the Regulation of Transmission Revenues Service. Standard Guidelines, Draft Decision", Available at http://www.accc.gov.au/content/index.phtml/itemId/345003/fromItemId/261500.

Acharyya, R. (1998), "Monopoly and product quality: separating or pooling menu?", Economic Letters, 61:187-94.

Afriat, S.N. (1972), "Efficiency estimation of production functions", International Economic Review, 58:568-98.

Ai, C. and D. Sappington (1998), "The Impact of State Incentive Regulation on the US Telecommunications Industry", University of Florida, December.

Aiyar, A.S.S. (1995), "Case for renegotiating Dabhol", The Economic Times, 17 July.

Ajodhia, V.S. (1999), "Betrouwbaarheidsanalyse van nieuwe distributienetten", MSc. Thesis, Delft University of Technology, Delft, January 1999.

Ajodhia, V.S., Hakvoort, R. and M. Van Gemert (2002), "Electricity Outage Cost Valuation: A survey", Proceedings of CEPSI 2002, Fukuoka.

Ajodhia, V.S. (2002), "Integrated Price and Reliability Regulation: The European Experience", Proceedings of the IEEE PES T&D Asia Pacific Conference, Yokohama.

Ajodhia, V.S., Franken, B. and C. Van der Lippe (2003), "Dutch Price and Quality Regulation System", Proceedings of the 26th IAEE International conference, Prague.

Ajodhia, V.S., Petrov, K. and G. Scarsi (2003), "Benchmarking and its Applications", Zeitschrift für Energiewirtschaft, 27:261-74.

Ajodhia, V.S., Petrov, K. and G. Scarsi (2004), "Quality, Regulation and Benchmarking. An application to electricity distribution networks", Zeitschrift für Energiewirtschaft, 28:107-20.

Ajodhia, V.S. and R.A. Hakvoort (2005), "Economic Regulation of Electricity Distribution Network Quality", Utilities Policy, 13:211-221.

Ajodhia, V.S., Kristiansen, T., Petrov K. and G. Scarsi (2005), "Total cost efficiency analysis for regulatory purposes: statement of the problem and two European case studies", 4th International Conference on Applied Infrastructure Research, October 2005, Berlin, Germany.

Ajodhia, V.S., Lo Schiavo, L. and R. Malaman (2006), "Regulation of Continuity of Supply in Italy – An Evaluation", Energy Policy (forthcoming).

Akerlof, G. (1970), "The Market for Lemons: Quality Uncertainty and the Market Mechanism", Quarterly Journal of Economics, 84:488-500.

Alexander, I. and T. Irwin (1996), "Price Caps, Rate-of-Return Regulation, and the Cost of Capital", Public Policy for the Private Sector No. 87, The World Bank, September.

Alexander, B. (1996), "How to Construct a Service Quality Index in Performance Based Ratemaking", The Electricity Journal, 9:46-53.

Allas, T. and K. Leslie (2001), "Sizing Power", The McKinsey Quarterly, 1:138-45.

Andersson, I. and L. Taylor (1986), "The social cost of unsupplied energy", Energy Economics, 8:139-46.

Arblaster, M. (1999), "Quality Service Monitoring", Utility Regulators Forum Discussion paper, Australia, October.

Armstrong, M., Cowan, S. and J. Vickers (1994), "Regulatory Reform. Economic Analysis and British Experience", MIT Press, Cambridge (MA).

Averch, H. and L. Johnson (1962), "Behavior of the Firm under Regulatory Constraint", American Economic Review, 52:1052-69.

Bailey, E. and R. Coleman (1971), "The effect of lagged regulation in the Averch-Johnson model", Bell Journal of Economics, 2:278-92.

Banerjee, A. (2003), "Does Incentive Regulation "Cause" Degradation of Retail Telephone Service Quality", Information Economics and Policy, 15:243-69.

Banker, R.D., Charnes, A. and W.W. Cooper (1984), "Some models for estimating technical and scale inefficiencies in data envelopment analysis", Management Science, 30:1078-92.

Baron, B. and R. Myerson (1982), "Regulating a Monopolist with Unknown Costs", Economica, 50:911-30.

Baumol, W.J. and A.K. Klevorick (1970), "Input Choices and Rate of Return Regulation: An Overview of the Discussion", Bell Journal of Economics, 1:162-90.

Bawa, V. and D. Sibley (1980), "Dynamic behaviour of a firm subject to stochastic regulatory review", International Economic Review, 21:627-42.

BBC (2002), "Who should run the railway system?", BBC Talking Point, 16 May 2002. Available at news.bbc.co.uk/1/hi/talking_point/1986551.stm.

Beenstock, M., Goldin, E. and Y. Haivosky (1998), "Response bias in a conjoint analysis of power outages", Energy Economics, 20:135-56.

Bental, B. and A. Ravid (1982), "A Simple Method for Evaluating the Marginal Cost of Unsupplied Energy", Bell Journal of Economics, 13:249-53.

Besanko, D., Donnenfeld, S. and L. White (1987), "Monopoly and quality distortion: Effects and Remedies", Quarterly Journal of Economics, 102:743-68.

Besanko, D., Donnenfeld, S. and L. White (1988), "The Multiproduct Firm, Quality Choice, and Regulation", Journal of Industrial Economics, 36:411-29.

Besanko, D. and S. Donnenfield (1988), "Rate of return regulation and product variety", Journal of Public Economics, 36:293-304.

Billinton, R., Wacker, G. and E. Wojczynski (1982), "Customer Damage Resulting from Electric Service Interruptions", Canadian Electrical Association, R&D project 907U 131, April.

Bloemhof, G.A. (2002), "Betrouwbaarheid van elektriciteitsnetten in Nederland in 2001", KEMA, Arnhem, Report nr. 40210010-TDC 02-27707A.

Boisseleau, F. (2004), "The role of power exchanges in the creation of a single European electricity market: Market design and market regulation", Delft University Press, Delft.

Boles, J.N. (1966), "Efficiency squared - Efficiency computation using efficiency indexes", Proceedings of the 39th annual meeting of the Western farm economics association.

Braeutigam, R.R. and J.C. Panzar (1993), "Effects of the change from rate-of-return regulation to price-cap regulation", American Economic Review, 83:191-98.

Brealy, R. and S. Myers (2000), "Principles of corporate finance", Irwin McGraw-Hill, Bostong.

Breyer, S. (1982), "Regulation and its reform", Harvard University Press, Cambridge (MA).

Brown, R. and J.R. Ochoa (1998), "Distribution system reliability: Default data and model validation", IEEE Transactions on Power Systems, 13:704-9.

Brown, R.E. (2002), "Electric Power Distribution Reliability", Marcel Dekker, Inc., New York.

Burns, G. and G. Gross (1990), "Value of service reliability", IEEE Transactions on Power Systems, 5:825-34.

Burns, P. and J. Davies (1998), "Regulatory incentives and capital efficiency in UK electricity distribution businesses", CRI Technical paper No. 12.

Cabral, L. (2000), "Introduction to industrial organisation", MIT Press, Cambridge (MA).

Caves, D., Herriges, J. and R. Windle (1990), "Customer Demand for Service Reliability in the Electric Power Industry: A Synthesis of the Outage Cost Literature", Bulletin of Economic Research, 2:79-119.

CEC (1997), "A survey of the implication to California of the August 10, 1996 Western States Power Outage", California Energy Commission, California.

CEER (2001), "Quality of Electricity Supply: Initial Benchmarking on Actual Levels, Standards and Regulatory Strategies", Council of European Energy Regulators.

CEER (2003), "Second Benchmarking Report on Quality of Electricity Supply", Council of European Energy Regulators.

CEPA (2003), "Background to work on assessing efficiency for the 2005 distribution price control review", Cambridge Economic Policy Associates.

Charnes, A., Cooper, W.W. and E. Rhodes (1978), "Measuring the efficiency of decision making units", European Journal of Operational Research, 2:429-44.

Chick, M. (2002), "Le tarif vert retrouve: The marginal cost concept and the pricing of electricity in Britain and France, 1945 - 1970", The Energy Journal, 23:97-116.

CIGRÉ (2004), "Power quality indices and objectives", Joint Working Group Cigré C4.07/Cired.

Clements, M. (2004), "Local Telephone Quality of Service: A Framework and Empirical Evidence", Telecommunications Policy, 28:413-26.

Coelli, T., Prasada Rao, D.S. and G.E. Battese (1998), "An introduction to efficiency and productivity analysis", Kluwer Academic Publishers, Boston.

CPB (2000), "Yardstick competition. Theory, design, and practice", CPB Working paper no. 133, Centraal Planbureau, The Hague, December.

Crew, M. and P. Kleindorfer (2002), "Regulatory Economics: Twenty Years of Progress?", Journal of Regulatory Economics, 21:5-22.

Cubbin, J. (2003), "Some more statistical properties of data envelopment analysis", mimeo, City University, London.

Davis, E. (1973), "A dynamic model of the regulated firm with a price adjustment mechanism", Bell Journal of Economics, 4:270-82.

De Nooij, M., Bijvoer, C. and C. Koopmans (2003), "The demand for supply security", Research Symposium European Electricity Markets, The Hague, September.

De Vries, L. (2004), "Securing the public interest in electricity generation markets. The myths of the invisible hand and the copper plate", Delft University of Technology, Delft.

Debreu, G. (1951), "The coefficient of resource utilisation", Econometrica, 19:273-92.

Demsetz, H. (1968), "Why regulate utilities?", Journal of Law and Economics, 11:56-65.

Dialynas, E.N., Megaloconomos, S.M. and V.C. Dali (2001), "Interruption cost analysis for the electrical power customers in Greece", Proceedings of CIRED 2001, Amsterdam.

Dixit, A. (1979), "Quality and quantity competition", Review of economic studies, 46:587-99.

Doane, M.J., Hartman, R.S. and C.K. Woo (1988), "Household preference for interruptible rate options and revealed value of service reliability", Energy Journal, 9:121-34.

Dorfman, R. and P.O. Steiner (1954), "Optimal advertising and optimal quality", American Economic Review, 44:826-36.

Douma, S. and H. Schreuder (1998), "Economic Approaches to Organizations", Prentice Hall International, Hertfordhire.

DTE (1999), "Price-cap regulering in de elektriciteitssector", DTE, Den Haag, available at www.dte.nl.

DTE (2000), "Guidelines for Price Cap Regulation in the Dutch Electricity Sector in the period from 2000 to 2003", DTE, Den Haag, available at www.dte.nl.

DTE (2002a), "Een Overzicht van de eerste Reguleringsperiode Regionale Netbedrijven Elektriciteit", DTE, Den Haag, available at www.dte.nl.

DTE (2002b), "Maatstafconcurrentie, Regionale Netbedrijven Elektriciteit", DTE, Den Haag, available at www.dte.nl.

DTE (2004a), "Besluit tot vaststelling van de methode van de korting ter bevordering van de doelmatige bedrijfsvoering ingevolge artikel 41, vierde lid, van de Elektriciteitswet 1998", DTE, Den Haag, available at www.dte.nl.

DTE (2004b), "Besluit tot vaststelling van de kwaliteitsterm ingevolge artikel 41a, eerste lid van de Elektriciteitswet 1998 voor de periode 2004 tot en met 2006", DTE, Den Haag, available at www.dte.nl.

DTE (2004c), "Consumenten en bedrijven tevreden met hoge betrouwbaarheid elektriciteitsnetten", DTE, Den Haag, available at www.dte.nl.

Dugan, R.C., McGranaghan, M.F. and H.W. Beaty (1996), "Electrical Power Systems Quality", McGraw-Hill, New York.

Energiened (2001), "Onvoorziene niet-beschikbaarheid in netten van 0 4 kV tot en met 150 kV in 2000", Energiened, Arnhem.

Farrel, M.J. (1957), "The measurement of productive efficiency", Journal of the Royal Statistical Society, 120:253-90.

Faucett, J. (1979), "Power shortage costs and effort to minimize: An example", EPRI report EA-1241, December.

Finsinger, J. and I. Vogelsang (1985), "Strategic management behaviour under reward structures in a planned economy", Quarterly Journal of Economics, 100:263-69.

Førsund, F. and S. Kittelsen (1998), "Productivity development of Norwegian electricity distribution utilities", Resource and Energy Economics, 20:207-224.

Fraser, R. (1994), "Price, Quality and Regulation: An Analysis of Price Capping and the Reliability of Electricity Supply", Energy Economics, 16:175-83.

Gabszewicz, J.J. and X.Y. Wauthy (2002), "Quality underprovision by a monopolist when quality is not costly", Economic Letters, 77:65-72.

Gal-Or, E. (1983), "Quality and quantity competition", Bell Journal of Economics, 14:590-600.

Gammelgard, M. and M. Larsson (2003), "The challenge of regulating natural monopolies in electrical distribution - Experiences from Sweden", Proceedings of CIRED 2003, Barcelona.

Gastaldo, E.M., Capeletto, G.J. and H.R. Bittencourt (2001), "Statistical correlations between indicators of power supply and consumers satisfaction", Proceedings of CIRED 2001, Amsterdam.

Gates, J., Billinton, R. and G. Wacker (1999), "Electric service reliability worth evaluation for government, institutions and office buildings", IEEE Transactions on Power Systems, 14:43-8.

Giannakis, D., Jamasb, T. and M. Pollitt (2005), "Benchmarking and incentive regulation of quality of service: an application to the UK electricity distribution networks", Energy Policy, 33:2256-71.

Glimer, R.W. and R.S. Mack (1983), "The cost of residential power outages", The Energy Journal, 4:55-74.

Gómez, T. and J. Rivier (2000), "Distribution and power quality regulation uder electricity competition. A comparative study", Proceedings of the 9th IEEE International Conference on Harmonics and Quality of Power ICHQP 2000. Orlando, Florida. October.

Green, R. and M. Rodriguez-Pardina (1999), "Resetting Price Controls for Privatized Utilities: A Manual for Regulators", World Bank Publications, Washington DC.

Grosfeld-Nir, A. and A. Tishler (1993), "A stochastic model for the measurement of electricity outage costs", Energy Journal, 14:157-174.

Hjalmarsson, L. and A. Veiderpas (1992), "Efficiency and ownership in Swedish electricity retail distribution", Journal of Productivity Analysis, 29:7-23.

Hougaard, J.L. (1994), "Produktivitetsanalyse af dansk elproduktion", Amternes og Kommunernes Forskningsinstitut, Copenhagen.

Hsu, G., Chang, P. and T. Chen (1994), "Various methods for estimating power outage costs: Some implications and results in Taiwan", Energy Policy, 22:69-74.

Huygen, A. (1999), "Regulering bij concurrentie. De Nederlandse elektriciteitssector", Universiteit Leiden, Leiden.

IEEE (1999), "IEEE trial-use guide for electric power distribution reliability indices", IEEE Std 1366-1998.

IPART (2001), "The Incorporation of Service Quality in the Regulation of Utility Prices", Research Paper No. 17, 2001, Independent Pricing and Regulatory Tribunal of New South Wales.

Jackson, A.F. and B. Salvage (1974), "Costs of electricity supply interruptions to industrial consumers (summary)", Proceedings of the IEEE, 121:1575-1576.

Jamasb, T. and M. Pollit (2000), "Benchmarking and Regulation: International Electricity Experience", Utilities Policy, 9:107-30.

Jamasb, T., Nillesen, P. and M. Pollitt (2003), "Gaming the Regulator: A Survey", The Electricity Journal, 16:68-80.

Jamasb, T. and M. Pollit (2003), "International benchmarking and regulation: an application to European electricity distribution utilities", Energy Policy, 31:1609-22.

Jamasb, T., Nillesen, P. and M. Pollitt (2004), "Strategic behaviour under regulatory benchmarking", Energy Economics, 26:825-43.

Jaromillo, P. and E. Skoknic (1973), "Costo social de las restricciones energia electrica", Officina de Planificacion, Chile, August 1973.

Jeunemaitre, A. and M. Matheu (2001), "Regulation of Network Utilities. The European Experience", Oxford University Press, New York.

Johnsonbough, R. (1993), "Discrete Mathematics", Macmillan, New York.

Joskow, P. (2005), "Incentive regulation in theory and practice: Electricity distribution and transmission networks", Working paper, Massachusetts Institute of Technology, 15 September.

Kahn, A. (1970), "The economics of regulation. Volumes 1 and 2", John Wiley and Sons, New York.

Kahn, E. (1997), "Determination of worth of reliability for better system planning and uses of system resources", Final report, ESAA/RSDC Project No. PN2420/95008, February.

Kahn, A. (2003), "The adequacy of prospective generation investments under price control mechanisms", The Electricity Journal, 15:37-46.

Kariuki, K.K. and R.N. Allan (1996a), "Assessment of customer outage cost due to service interruptions: Residential sector", Proc. IEE - Gener. Transm. Distrib., 143:163-70.

Kariuki, K.K. and R.N. Allan (1996b), "Evaluation of reliability worth and value of lost load", Proc. IEE - Gener. Transm. Distrib., 143:171-180.

KEMA (2002), "Report on Regulation of Asset Management", Report to DTE, November.

KEMA (2003), "Wensstromen", Rapport aan het Ministerie van Economische Zaken in het kader van PREGO, Arnhem, Netherlands.

KEMA (2004), "Price cap regulation in electricity distribution", mimeo, KEMA Consulting, Bonn.

Khator, S.K. and L.C. Leung (1997), "Power Distribution Planning: A review of models and issues", IEEE Transactions on Power Systems, 12:1151-59.

Kidokoro, Y. (2002), "The Effects of Regulatory Reform on Quality", Journal of Japanese and International Economies, 16:135-46.

Kim, J. and J. Kim (1996), "Quality choice of multi-product monopolist and spill-over effects", Economic Letters, 52:345-52.

Klaassen, K.B., Van Peppen, J.C.L. and A. Bossche (1988), "Bedrijfszekerheidstechniek Theorie en Praktijk", Delftse Uitgeversmaatschappij, Delft.

Klevorick, A. (1973), "The behaviour of a firm subject to stochastic regulatory review", Bell Journal of Economics, 4:57-88.

Kling, W.L. (1994), "Planning en bedrijfsvoering van Elektriciteitsvoorzieningssystemen", Delft University of Technology, Delft.

Knops, H. (2005), "PhD thesis", Delft University of Technology, Delft (forthcoming).

Koopmans, T.C. (1951), "An analysis of production as an efficient combination of activities", in: Koopmans, T C (Ed.) Activity analysis of production and allocation, Cowles commission for research in economics, Monograph No. 13, Wiley, New York.

Korhonen, P. and M. Syrjänen (2002), "Evaluation of cost efficiency in Finnish Electricity Distribution", Working paper, Helsinki School of Economics.

Krohm, G.C. (1978), "A survey of disruption and consumer costs resulting from major residential power outage", Argonne National Laboratory, Report no. ANL/EES-TM-29 and 30, Argonne, IL, USA, June.

Kumbhakar, S. and L. Hjalmarsson (1998), "Relative performance of public and private ownership under yardstick competition: Electricity retail distribution", European Economic Review, 42:97-122.

Laffont, J.J. and J. Tirole (1986), "Using cost observation to regulate firms", Journal of Political Economy, 94:614-641.

Laffont, J.J. and J. Tirole (1993), "A Theory of Incentives in Regulation and Procurement", MIT Press, Cambridge (MA).

Lambertini, L. (1998), "Does Monopoly Undersupply Product Quality?", Dipartimento di Scienze Economiche, Univrsita degli Studi di Bologna, April.

Lancaster, K. (1975), "Socially optimal product differentiation", American Economic Review, 65:567-85.

Langset, T., Trengereid, F., Samdal K. and J. Heggset (2001), "Quality Dependent Revenue Caps - A Model for Quality of Supply Regulation", Proceedings of CIRED 2001, Amsterdam.

Larsson, M.B.O. (2003), "The network performance assessment model - Regulation with a reference network", Proceedings of the Elforsk Market Design Conference, Stockholm, 17 June.

Lee, D.T.R. and B.J. Schachter (1980), "Two algorithm for constructing a Delaunay triangulation", International Journal of Computer and Information Science, 9:219-42.

Lehtonen, M. and B. Lemstrom (1995), "Comparison of the methods for assessing the customers outage costs", Proceedings of EMPD '95.

Leibenstein, H. (1966), "Allocative Efficiency versus X-Efficiency", American Economic Review, 56:392-415.

Leland, H. (1979), "Quacks, lemons and licensing: A theory of minimum quality standards", Journal of Political Economy, 87:1328-46.

Levhari, D. and Y. Peles (1973), "Market structure, quality and durability", Bell Journal of Economics, 4:235-48.

Liston, C. (1993), "Price-Cap Versus Rate-of-Return Regulation", Journal of Regulatory Economics, 5:25-48.

Littlechild, S. (1983), "Regulation of British Telecommunications' profitability", Department of Industry, London.

Lo, F., Chien, C. and J.T. Lin (2001), "A DEA Study to Evaluate the Relative Efficiency and Investigate the District Reorganization of the Taiwan Power Company", IEEE Transactions On Power Systems, 16:170-8.

Lo Schiavo, L. and R. Malaman (2003), "Regulation and Incentives for Improving Continuity of Supply: The Experience of Italy and a Comparison with other EU Countries", Proceedings of CIGRE/IEEE Symposium on Quality and Security of Supply, Montreal.

Loeb, M. and W.A. Magat (1979), "A Decentralized Method of Utility Regulation", Journal of Law and Economics, 22:399-404.

Logan, J., Mason, R. and R. Reynolds (1989), "Efficient regulation with little information: Reality or limit?", International Economic Review, 30:851-62.

Lolander, D. (1948), "Memorandum concerning estimation of the cost of the national economy in the event of a loss of supply", Vattenfall.

London Economics (1999), "Efficiency and benchmarking study of New South Wales distribution businesses", Report commissioned by the Independent Pricing and Regulatory Tribunal (IPART) of New South Wales.

Meeuwsen, J.J. (1998a), "Handleiding Distrel", Technische Universiteit Delft, Faculteit Informatietechnologie en Systemen, Afdeling Elektrotechniek.

Meeuwsen, J.J. (1998b), "Reliability evaluation of electric transmission and distribution systems", Technische Universiteit Delft, Delft.

Mikkers, M.C. and V. Shestalova (2001), "Yardstick Competition and Reliability of Supply in Public Utilities", Working Paper, Tilburg University/Office For Energy Regulation, Netherlands.

Modern Manufacturing (1969), "Power survey - problems and progress", Modern Manufacturing, July 1969.

Munasinghe, M. (1979), "The Economics of Power System Reliability and Planning. Theory and Case Study", Johns Hopkins University Press, Baltimore.

Munasinghe, M. (1980), "The Costs Incurred by Residential Electricity Consumers Due to Power Failures", Journal of Consumer Research, 6:361-369.

Munasinghe, M. (1981), "Optimal Electricity Supply. Reliability, pricing and system planning", Energy Economics, 3:140-52.

Munasinghe, M. and J.J. Warford (1982), "Electricity Pricing", Johns Hopkins University Press, Baltimore.

Munasinghe, M. (1984), "Engineering-Economic analysis of electric power systems", Proceedings of the IEEE, 72:424-61.

Munasinghe, M. (1990), "Energy Analysis and Policy", Butterworth-Heinemann Press, London.

Mussa, M. and S. Rosen (1978), "Monopoly and product quality", Journal of Economic Theory, 18:301-17.

NARUC (1997), "Performance-Based Regulation in a Restructured Electric Industry", Report for NARUC by Synapse Energy Economics, November 8.

Newbery, D. (1997), "Privatisation and liberalisation of network utilities", European Economic Review, 41:357-83.

Newbery, D. (1999), "Privatisation, restructuring and regulation of network utilities", MIT Press, Cambridge (MA).

Ofgem (2001), "The Information and Incentives Project. Developing the incentive scheme. Final proposals", Ofgem, London, England.

Ofgem (2002), "Asset Risk Management in electricity and gas networks – Conclusions on framework and proposals for the survey", Ofgem, London, England.

Pahwa, A., Feng, X. and D. Lubkeman (2002), "Performance evaluation of electric distribution utilities based on data envelopment analysis", IEEE Transactions on Power Systems, 17:400-5.

Pandey, M. and R. Billinton (1999), "Reliability worth assessment in a developing country-residential survey results", IEEE Transactions on Power Systems, 14:1226-31.

Pardina, M.R. and M. Rossi (2000), "Technical Change and Catching-up: The Electricity Distribution Sector in South America", Working Paper No. 11. Centro de Estudios Económicos de la Regulación, Departamento de Economía y Finanzas, Universidad Argentina de la Empresa, November.

Parikh, J., Chattopadhyay, D. and U. Nandapurkar (1995), "Simulation of national grid operation in India", Utilities Policy, 5:65-74.

Peco, J. and T. Gómez (2000), "A Model for Planning Large Scale MV Networks considering Uncertainty in Reliability Modeling", Proceedings of PMAPS 2000, Portugal.

Peltzman, S. (1976), "Toward a More General Theory of Regulation", Journal of Law and Economics, 19:211-40.

Peltzman, S. (1989), "The economic theory of regulation after a decade of deregulation", Brookings papers on economic activity: Microeconomics, 1-59.

Peréz-Arriaga, I.J. (2004), "Personal communications with Prof. Peréz-Arriaga ", June.

Pollit, M. (1995), "Ownership and performance in electric utilities: The international evidence on privatisation and efficiency", Oxford University Press, New York.

Posner, R.A. (1971), "Taxation by regulation", Bell Journal of Economics, 2:22-50.

Posner, R.A. (1974), "Theories of economic regulation", Bell Journal of Economics, 5:335-58.

Raza, A. (2003), "Establising a regulatory framework for quality of service incentives", Regulateri. Quartely review of regulatory developments, Issue 16, March.

Resende, M. (2002), "Relative efficiency measurement and prospects for yardstick competition in Brazilian electricity distribution ", Energy Policy, 30:637-47.

Rivier, J. and T. Gómez (2000), "A conceptual framework for power quality regulation", Proceedings of the 9th IEEE International Conference on Harmonics and Quality of Power ICHQP 2000. Orlando, Florida. October.

Rivier, J. and T. Gómez (2003), "Critical analysis of Spanish power quality regulation design", Proceedings of the Market Design Conference 2003, pp. 29-38. Stockholm, Sweden. June 16-17.

Rivier, J., Gómez, T., De la Fuente León J.I. and J. Román (1999), "Power quality regulation under the new regulatory framework of distribution systems.", Proceedings of the 13th PSCC, Vol. 2, pp. 891-897, Trondheim, Norway, June.

Rossi, M.A. and C.A. Ruzzier (2000), "On the regulatory application of efficiency measures", Utilities Policy, 9:81-92.

Rovizzi, L. and D. Thompson (1995), "The Regulation of Product Quality in the Public Utilities", in M. Bishop, J. Kay, C. Mayer (Eds.) The Regulatory Challenge, Oxford University Press, Oxford.

Roycroft, T. and M. Garcia-Murillo (2000), "Trouble Reports as an Indicator of Service Quality: The Influence of Competition", Telecommunications Policy, 24:947-67.

Rudnick, H. and R. Raineri (1997), "Chilean distribution tariffs: Incentive Regulation", in: (De)Regulation and competition: The electricity industry in Chile. Iiades-Georgetown Unversity pp. 223-257.

Rudnick, H. and J.A. Donoso (2000), "Integration of price cap and yardstick competition schemes in electrical distribution regulation", IEEE Transactions on Power Systems, 15:1428-33.

Sanghvi, A. (1982), "Economic Costs of Electricity Supply Interruptions. US and Foreign Experience", Energy Economics, 4:180-98.

Sappington, D. and D. Sibley (1988), "Regulating without Cost Information: The Incremental Surplus Subsidy Scheme", International Economic Review, 29:297-306.

Sappington, D. and D. Weisman (1996), "Revenue Sharing in Incentive Regulation Plans", Information Economics and Policy, 8:229-48.

Sappington, D., Pfeifenberger, J., Hanser P. and G. Basheda (2001), "The State of Performance-Based Regulation in the US Electric Utility Industry", The Electricity Journal, 14:71-79.

Sappington, D. (2003), "The Impact of Incentive Regulation on Retail Telephone Service Quality in the United States", Review of Network Economics, 2:355-75.

Sarkar, A. and R.M. Shrestha (1996), "Integrated analysis of customer value of generation system reliability in India", Energy Sources, 18:569-83.

Scarsi, G. (1999), "Local electricity distribution in Italy: Comparative efficiency analysis and methodological cross-checking", Fondazione ENI Enrico Mattei (FEEM), Milan, Working Paper no. 16/1999.

SCI (1978), "Impact assessment of the 1977 New York City Blackout", System Control Inc (SCI), Arlington VA, July.

Seiford, L. and R. Thrall (1990), "Recent developments in DEA: The mathematical programming approach to frontier analysis", Journal of Econometrics, 46:7-38.

SEO (2004), "Op prijs gesteld, maar ook op kwaliteit: de prijs van stroomonderbrekingen – op zoek naar phi", SEO Economisch Onderzoek, Amsterdam, Publication number 726.

Shalaan, A.M. (1989), "Electric service interruptions impact and cost estimation", Electr Cigre 127, December.

Sheppard, H.J. (1967), "The economics of reliability of supply distribution", The Economics of Reliability of Supply, IEEE Conference publication, 34:248-266.

Sheshinski, E. (1976), "Price, Quality and Quantity Regulation in Monopoly Situations", Economica, 43:127-37.

Sheshinski, E. (2003), "Personal communications with Prof. Eytan Sheshinski", June.

Shipley, R.B., Patton, A.D. and J.S. Denison (1972), "Power Reliability – Cost vs Worth", IEEE Winter Power Meeting, Paper No. T-72.

Shleifer, A. (1985), "A Theory of Yardstick Competition", Rand Journal of Economics, 16:314-27.

Sibley, H. (2002), "Loss Aversion, Price and Quality", University of Tasmania, Australia, October.

Smith, W. (1997), "Utility Regulators – The Independence Debate", The Private Sector in Infrastructure: Strategy, Regulation, and Risk. The World Bank Press, Washington, DC.

Spence, A.M. (1975), "Monopoly, Quality and Regulation", Bell Journal of Economics, 6:417-29.

Steetskamp, I. and A. Van Wijk (1994), "Stroomloos: Kwetsbaarheid van de samenleving: gevolgen van verstoringen in de electriciteitsvoorziening", Rathenau Instituut, the Netherlands.

Stigler, G. (1971), "The theory of economic regulation", Bell Journal of Economics, 2:3-21.

Stiglitz, J. and C. Walsh (2002), "Principle of macroeconomics", Norton, New York.

Sullivan, M. and M. Sheehan (2000), "Observed Changes in Residential and Commercial Customer Interruption Costs in The Pacific Northwest between 1989 and 1999", IEEE PES Summer Meeting, 4:2290-3.

Sumicsid (2003a), "Norm models", Report to NVE, February.

Sumicsid (2003b), "Benchmarking for Regulation", Report to NVE, August.

Swan, P. (1970), "Durability of consumer goods", American Economic Review, 60:889-94.

Swedish Joint Commission (1969), "Cost of the interruption in Electricity Supply", Stockholm, September.

Taiwan Power Company (1980), "Evaluation of generating capacity reserve for Taiwan Power Company", in: M. Munasinghe, The Economics of Power System Reliability and Planning, Johns Hopkins University Press, 1979.

Tangerås, T. (2002), "Regulation of Cost and Quality under Yardstick Competition", Working paper no. 573, The Research Institute of Industrial Economics, Stockholm.

Tavanir (1995), "Electricity interruption cost in Iran", Power Planning Bureau, Tavanir company, Ministry of Energy, Volume 2-6.

Telson, M. (1975), "The Economics of Alternative Levels of Reliability for Electric Power Generation Systems", Bell Journal of Economics, 6:679-94.

Ter-Martirosyan, A. (2003), "The Effects of Incentive Regulation on Quality of Service in Electricity Markets", Dpt. Of Economics, George Washington University, Working Paper, March.

Tersztyanszky, T. (2003), "Personal communications with dr. Tibor Tersztyanszky", Hungarian Energy Office (HEO), Budapest, May.

Train, K.E. (1991), "Optimal Regulation", MIT Press, Cambridge (MA).

Trengereid, F. (2003), "Quality of supply regulation in Norway", Norwegian water resources and energy directorate (NVE).

Turner, A.J. (1977), "The costs to consumers of interruptions to the electricity supply", mimeo, New Zealand Electricity, May.

Turvey, R. (2001), "What are marginal costs and how to estimate them", TechnicalPaper 13 by Centre for the study of Regulated Industries.

TZH (2000), "Presentation by TZH to DTE", The Hague, September.

UNIPEDE (1972), "Quality of service from the consumer's point of view", UNIPEDE report No 60/D1, Paris.

Upadhyay, A.K. (1996), "Power at any price", The Economic Times, 1 January.

Varian, H. (1993), "Economic Incentives in Software Design", Computational Economics, 6:201-17.

Vickers, J. and G. Yarrow (1988), "Privatisation: An Economic Analysis", MIT Press, Cambridge (MA).

Vogelsang, I. and J. Finsinger (1979), "A regulatory adjustment process for optimal pricing by multiproduct monopoly firms", Bell Journal of Economics, 10:157-71.

Warren, C.A., Ammon, R. and G. Welch (1999), "A survey of distribution reliability measurement practices in the US", IEEE Transactions on Power Delivery, 14:250-7.

Weisman, D. (2002), "Price, Regulation and Quality", mimeo, Kansas State University, July.

Weyman-Jones, T. (1991), "Productive efficiency in a regulated industry. The area electricity boards of England and Wales", Energy Economics, 13:116-22.

Weyman-Jones, T. (1995), "Problems of Yardstick Regulation in Electricity Distribution", in M. Bishop, J. Kay, C. Mayer (Eds.) The Regulatory Challenge, Oxford University Press, Oxford.

Whitaker, R. (2002), "Personal communications with Robert Whitaker", Office of Regulatory Economics, New York State Public Service Commission, January.

White, L. (1977), "Market structure and product varieties", American Economic Review, 67:179-82.

Wijayatunga, P. and M. Jayalath (2004), "Assessment of economic impact of electricity supply interruptions in the Sri Lanka industrial sector", Energy Conversion and Management, 45:235-47.

Williamson, O. (1976), "Franchise bidding for natural monopolies- In general and with respect to CATV", Bell Journal of Economics, 7:73-104.

Williamson, B. (2001), "Incentives For Service Quality: Getting the Framework Right", The Electricity Journal, 14:62-70.

Willig, R. (1976), "Consumer's Surplus Without Apology", American Economic Review, 66:589-97.

Willis, H.L. (1997), "Power distribution planning reference book", Marcel Dekker, Inc., New York.

Woo, C.K. and R.L. Pupps (1992), "Cost of service disruptions to electricity customers", Energy, 17:109-26.

Young, A. (1987), "The value of electricity reliability", Report to electricity distribution division, New Zealand Ministry of Energy, January.

List of Symbols

Symbols	
AC	Average costs
AE	Allocative efficiency
CP	Consumer surplus
CPI	Consumer Price Index
D	Demand
Dep	Depreciation
Du	State duration
EE	Economic efficiency
ENS	Energy not supplied
F	Number of failures
FC	Fixed costs
Fr	State frequency
I	Electrical current
ICENS	Interruption costs per kWh
Inv	Investment
K	Capital
L	Labour
N	Number of firms (section 5.2)
N	Number of consumers (section 5.3)
N	Number of system states (chapter 6)
LF	Load factor

NB	Net benefit
NC	Network costs
P	Inverse demand
PD	Peak demand
PPI	Producer Price Index
Pr	State probability
Q	Output
R	Reliability
R*	Expected reliability
RAB	Regulatory asset base
Rev	Revenue
RPI	Retail Price Index
S	Parameter Sensitivity
TB	Total benefit
TC	Total costs
TE	Technical efficiency
W	Welfare
WACC	Weighted Average Costs of Capital
X	Productivity improvement factor
a	Asset type
c	Costs
capex	Capital expenditures
d	Interruption duration
f	Interruption frequency
ic	Interruption costs
losses	Electrical losses
n	Duration of the regulatory period
nc	Network costs

opex	Operational expenditures
p	Price
p_L	Price of losses
q	Quality
r	Interest rate
ror	Rate-of-return
s	Inverse quality
sotex	Social costs
t	Time
tr	Monetary transfer
$\mathbf{u}$	Output weight
$\mathbf{v}$	Input weight
v	General variable
w	Price for labour
x	Quantity
$\mathbf{X}$	Input factor
$\mathbf{Y}$	Output factors
Z	Environmental variable
α	Fixed monetary transfer
β	Fraction of costs borne
η	Productivity level
θ	Efficiency score
λ	Failure rate (chapter 6)
λ	Frontier projection coefficient (chapter 5)
μ	Repair time
π	Profit
τ	Time
ϕ	Quality incentive

φ	Willingness to pay for quality
Ω	Ohmic resistance
Indices	
H	High quality
L	Low quality
M	Minimum quality
a	Asset type
c	Connection
d	DOWN-state
i	Consumer
j	Firm
k	Component
l	System state
m	System state
u	UP-state

Annex I. Quality Indicators

Customer-based Indicators

Customer-based indicators use information on the number of interrupted customers and the duration that these customers were interrupted. The main advantage of these indicators is that they consider reliability from the customer perspective. This makes them quite popular not only amongst firms but in particular, amongst regulators who prefer relatively simple and easy to interpret measures of reliability performance. A survey conducted by the IEEE revealed that the most commonly used indicators are SAIFI, SAIDI, CAIDI, and ASAI (Warren et al. 1999).

- SAIFI measures the probability that a customer will experience an outage. It is calculated by dividing the number of customer interruptions by the total number of customers served. The number of customer interruptions is the total number of interrupted customers for each interruption. This is typically measured over the period of a calendar year.
- SAIDI provides a measure for the average time that customers are interrupted. It is calculated by dividing the total customer interruption duration by the total number of customers. The customer interruption duration is defined as the aggregated time that all customers were interrupted. SAIDI is also known as Customer Minutes Lost (CML).
- CAIDI is defined as SAIDI divided by SAIFI and is a measure for the average time required restoring service to the average customer per interruption. It is calculated by dividing the total interruption duration by the total number of interruptions.

From the three above-mentioned indicators, SAIDI (or CML) is most commonly used by distribution firms as well as by regulators. For example, in the US over 80 percent of utilities use this indicator (Warren et al. 1999) while there is usually also a regulatory requirement to measure and report SAIDI (CEER 2003). In formulaic terms, the three above indicators are calculated as follows:

System Average Interruption Frequency Index

$$\text{SAIFI} = \frac{\text{Total Number of Customer Interruptions}}{\text{Total Number of Customers Served}} = \text{CI} / 100 \qquad \text{/yr}$$

System Average Interruption Duration Index

$$\text{SAIDI} = \frac{\text{Customer Interruption Durations}}{\text{Total Number of Customers Served}} = \text{CML} \qquad \text{hr/yr}$$

Customer Average Interruption Duration Index

$$\text{CAIDI} = \frac{\text{Customer Interruption Durations}}{\text{Total Number of Customer Interruptions}} \qquad \text{hr}$$

To better understand the SAIFI, CAIDI, and SAIDI indicators, it is helpful to explain the concept of customer interruption duration (CID). For an individual customer, the interruption duration is given by the time power is interrupted. CID is the cumulated duration of interruption for all customers and for all customers in a given period (e.g. a year). In case of an interruption, the interruption duration per customer may differ. The reason for this is that, usually, service is restored to customers in phases (see Box I-1 for an example).

Box I-1. Example of restoring electricity supply in case of an interruption

Consider the simple distribution network in Figure I-1. This network consists of six consumers, which are supplied by the main station. Supply takes place via a loop, which is divided, into two parts. The two outgoing feeders of the main station are protected: In case of a fault in the loop, protection automatically disconnects the affected feeder i.e. only half of the consumers are interrupted. Assume now that a fault occurs between loads 2 and 3. This leads all customers connected to the first part of the loop (1, 2, and 3) to be interrupted. After the interruption occurs, the first step will be to locate the fault. Then, the faulted part of the cable, between loads 2 and 3, will be isolated. After that, supply will be restored. This may be done by re-energising the first feeder (which reconnects loads 1 and 2) and then by closing the open loop (which reconnects load 3). Alternatively, these steps may be taken in reverse order. Irrespective, as may be observed, there will be differences in interruption duration time between loads 1 and 2, and load 3, respectively.

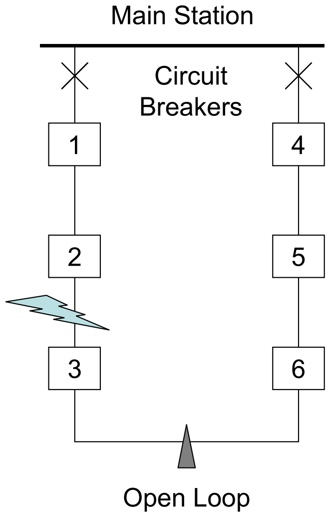


Figure I-1. Schematic overview of a simple distribution network.

In principle, the SAIFI, SAIDI, and CAIDI indicators are not a measure for reliability but rather for *unreliability* as they relate to customers interrupted. An indicator that measures reliability but which is less commonly used is ASAI. This represents reliability as a number between zero (no supply at all) and hundred percent (no interruptions at all).

- ASAI is the customer-weighted availability of the system and provides the same information as SAIDI. This index represents the fraction of time (often a percentage) that a customer has power during either one year or the defined reporting period. Higher ASAI values reflect higher levels of reliability. ASAI is calculated as follows:

Average Service Availability Index

$$\text{ASAI}= \frac{\text{Customer Hours Service Availability}}{\text{Customer Hours Service Demand}} \quad \%$$

Sometimes, interruptions are not sustained but only temporary. This means that the duration of the interruption is limited to a "short" period, after which restoration takes place very quickly without the explicit intervention of the utility e.g. through automatic switching. Usually, a threshold value is chosen for the interruption duration to classify it as either momentary or sustained. If the momentary interruption lasts longer than some predefined period, it becomes a sustained interruption. A common threshold value is three minutes, which implies that all interruptions shorter than three minutes will be classified as momentary and all other interruptions as sustained. Usually, momentary interruptions are not included in the reliability indicators but are instead recorded through a separate index, MAIFI.

- MAIFI is similar to SAIFI, with the difference that it measures only momentary interruptions. SAIFI, on the other hand, measures sustained interruptions although - in some cases - SAIFI may include both sustained and momentary interruptions.

Momentary Average Interruption Frequency Index

$$\text{MAIFI}= \frac{\text{Total Number of Customer Momentary Interruptions}}{\text{Total Number of Customers Served}} \text{ /yr}$$

One of the important debating points between regulators and firms is whether momentary interruptions should also be considered in the regulatory framework. Historically, utilities have tended to exclude momentary interruptions from published reliability statistics. However, given the increased sensitivity of customer loads to brief interruptions, regulators have started questioning this practice and are now more inclined to include all interruptions in the performance reporting and – if applicable – use them as part of quality incentive mechanisms.

Some less commonly used reliability indicators are not based on the total number of customers served, but on the number of customers that experienced one or more interruptions in the given time period. These indicators are more difficult to calculate, as one would need to distinguish between customers with and without interruptions.

- CAIFI is a hybrid of SAIFI and is calculated in the same way, with the exception that the normalisation is based only on customers who have actually experienced an interruption.
- CTAIDI is a hybrid of CAIDI, and is calculated in the same way with the exception that the normalisation is based only on customers who have actually experienced an interruption.

Customer Average Interruption Frequency Index

$$\text{CAIFI}= \frac{\text{Total Number of Customer Interruptions}}{\text{Customers Experiencing 1 or more interruptions}} \quad /\text{yr}$$

Customer Total Average Interruption Duration Index

$$\text{CTAIDI}= \frac{\text{Customer Interruption Durations}}{\text{Customers Experiencing 1 or more interruptions}} \quad \text{hr/yr}$$

On the surface, CAIFI may seem similar to SAIFI. However, the two are fundamentally different. One large difference is that the lowest possible value for SAIFI is zero, while the lowest possible value for CAIFI is one. Reducing the number of interruptions that a customer experiences from, say, two to one will improve SAIFI (as, on average, customers are experiencing fewer interruptions). However, note that reducing the number of interruptions that this same customer experiences from one to zero will make CAIFI *worse*. This is because the denominator in the formula (the number of customers experiencing an interruption) will decrease; hence, the value for CAIFI will increase. Thus, improvements in CAIFI do not necessarily correspond to improvements in reliability, which makes the statistic flawed. Similar problems occur also with CTAIDI, which can be improved by increasing the number of customers that experience just one interruption and is therefore prone to some degree of manipulation.

Table I-1. Example of difference in calculations between SAIFI and CAIFI.

	Nr. of interruptions per customer					
	Cust. A	Cust. B	Cust. C	Cust. D	SAIFI	CAIFI
Case 1:	2	0	1	3	(2+0+1+3)/**4**=1.5	(2+0+1+3)/**3**=1.5
Case 2:	1	0	1	3	(1+0+1+3)/**4**=1.25	(1+0+1+3)/**3**=1.25
Case 3:	0	0	1	3	(0+0+1+3)/**4**=1.0	(0+0+1+3)/**2**=1.0

The previously discussed reliability indicators reflect *average* system reliability. These measures may not necessarily reflect consumer satisfaction since few customers may actually experience average reliability. Some consumers may experience very high reliability while others may be confronted with a large number of interruptions. A different statistic, CEMI_n, takes into account the distribution in the frequency of interruptions across customers.

- CEMI_n aims to examine variations in customer reliability and to identify the number of customers with relatively poor reliability. The subscript *n* defines a threshold level of reliability that a customer must exceed before being counted.

Customers Experiencing Multiple Interruptions

$$\text{CEMI}_n = \frac{\text{Cust. Experiencing More than } n \text{ Interruptions}}{\text{Total Number of Customers Served}} \quad /\text{yr}$$

CEMI_n can provide useful information about differences in reliability levels throughout the system. In particular, it can help quantify the number of worst served customers and is therefore interesting from a regulatory point of view. However, due to the heavy data requirements in calculating this indicator, it is not commonly used. Also, it is less straightforward to interpret by the general public compared to more traditional indicators.

Load-based Indicators

Customer-based indicators weighs each customer in the same way. They do not distinguish between differences that may exist between customers in terms of size, demand, energy consumption, extent of interruption damage, etc. By treating each customer equally, these indicators do not reflect the differences in impact caused to different customers. For example, the costs experienced by a large industrial plant may well be equal to those experienced by a thousand households. However, this effect will not be reflected in a customer-based indicator, as the industrial client and the household will be weighted equally. To make indicators reflect more closely the differences in interruption impact and costs, one could measure reliability indicators separately for each customer group. This, however, is typically costly to do in the face of data and measurement limitations.

Load and energy-based indicators can help (partially) capture the heterogeneity between customers. The general idea is that the larger consumers, as approximated by the size of connected capacity or consumer electricity, generally incur higher costs than smaller consumers do. By weighting interruptions on the basis of interrupted load or non-supplied energy (instead of the customer numbers), the indicator can more closely reflect the impact of interruptions.

- ASIFI corresponds to SAIFI but - instead of the number of interrupted customers – it makes use of the amount of capacity disconnected during the interruption. This disconnected capacity is then normalised by the total served capacity by the network.
- ASIDI correspond to SAIDI but similarly to ASIFI, it uses connected capacity instead of customer numbers for the purpose of weighting and normalisation.

ASIFI and ASIDI are also known as NIEPI and TIEPI, (in Spain), or as FMIK and TTIK (in Latin America). They are calculated as follows:

Average System Interruption Frequency Index

$$\text{ASIFI} = \frac{\text{Connected kVA Interrupted}}{\text{Total Connected kVA Served}} \qquad \text{/yr}$$

Average System Interruption Duration Index

$$\text{ASIDI} = \frac{\text{Connect kVA Hours Interrupted}}{\text{Total Connected kVA Served}} \qquad \text{hr/yr}$$

It is worth mentioning that ASIFI and ASIDI are actually one of the oldest indicators used in the electricity industry. The reason for this is historical. In the past, utilities knew the size of their distribution transformers but did not know how many customers were connected to each transformer. In the case of an interruption, it was relatively easy to determine the amount in kVA of interrupted transformer capacity. However, the use of transformer capacity as a proxy for disconnected load overestimates the statistic, as transformer capacity in general will be higher than actual installed capacity served by the transformer. In principle, the actually connected capacity should be used rather than the nominal capacity of the transformer. Today, the availability of customer information systems makes it much easier to determine the number of customers connected to each transformer, and subsequently to calculate customer-based indicators.

Energy-based Indicators

Energy-based indicators are closely related to load-based ones. These indicators consider the amount of energy not supplied because of interruptions, which is typically normalised by the number of connected customers. In formulaic terms:

Energy Not Supplied

$$\text{ENS} = \frac{\text{Energy Not Supplied}}{\text{Total Number of Customers Served}} \qquad \text{kWh/year}$$

ENS is strongly correlated with the costs that consumers experience during interruptions. Usually, units of interruption costs are expressed in terms of cost per kWh, thus making the ENS indicator proportional to these costs. This feature is for example used by the Norwegian regulator (NVE), who adjusts the level of allowed revenues on the basis of the difference between actual and expected interruption costs. These interruption costs are defined as the amount of energy not supplied, multiplied by the “price” for each kWh not supplied and differentiated by consumer class.

As noted earlier, there are some difficulties attached to the measurement of interrupted capacity and, usually, transformer capacity is used as a proxy for this. Similar problems apply to ENS. For larger customers, there will usually be metering installed and thus ENS can be quite accurately measured. In Norway for example, 50 percent of energy distributed is measured directly. The remaining 50 percent needs estimating via load profiles to derive information about ENS. The idea is that during each period (e.g. an hour), each type of customer will have a certain level of consumption. If these load profiles are known for each type of customer and there is information about the number of each customer type per transformer, then ENS can subsequently be estimated. Such an approach is the one used in Norway.

A final note to be made is that ENS does not necessarily reflect the *actual* amount of energy that customers fail to consume as a result of being interrupted. Generally, one may note some “catching-up” effects immediately after an interruption as customers may temporarily use more energy to partially make up for the period during which they were interrupted.

International Experiences

From the previous section, a wide range of indicators has emerged. However, as Table I-2 shows, there is a strong preference by regulators to use customer-based indicators such as SAIDI or SAIFI. This strong preference for customer-based indicators may be attributed to different reasons:

- It is in line with practices within the utility industry itself.
- These indicators provide a measure of the performance level that is actually supplied to consumers. This is what regulators are mostly concerned with.
- These indicators are relatively easy to interpret and to communicate to the public, who generally do not have sufficient technical knowledge to fully understand more complex indicators.
- These indicators require relatively little data, which makes the problem of auditing these data less difficult for the regulator.

Only in a few instances are load and/or energy-based indicators used. Particularly in Spain, the regulator makes use of ASIFI and ASIDI (locally, NIEPI and TIEPI respectively). For the time being, as far as known, Norway is the only country where the regulator actively uses energy-based indicators.

Table I-2. International survey of reliability indicators used by regulators.

Country	Basis for Quality Indicators		
	Customers	**Load**	**Energy**
Australia (Victoria)	SAIFI, SAIDI, CAIDI		
Belize	SAIDI, SAIFI		
France	SAIFI		
Italy	SAIDI		
The Netherlands	SAIFI, CAIDI		
Norway			ENS
Portugal	SAIFI, SAIDI	TIEPI	
Spain		TIEPI, NIEPI	
Sweden	SAIFI, CAIDI		
United Kingdom	CML, CI		

Annex II. Benchmarking Techniques

All benchmarking techniques have in common that they try to estimate the frontier i.e. most productive firms using sample data. According to Coelli et al. (1998), this data may involve observations on a number of firms in a particular time interval (cross-sectional data); aggregate industry-level data observed over a number of time intervals (time-series data); or observations of a number of firms in a number of time intervals (panel data). Sample analysis is used by partial methods who consider only a single input and output combination at the time, and by total methods, which consider multiple input and output combinations at the same time. Additionally, norm models make use of an artificially constructed sample e.g. by means of optimising or simulation techniques. The benchmarking techniques are now briefly discussed.

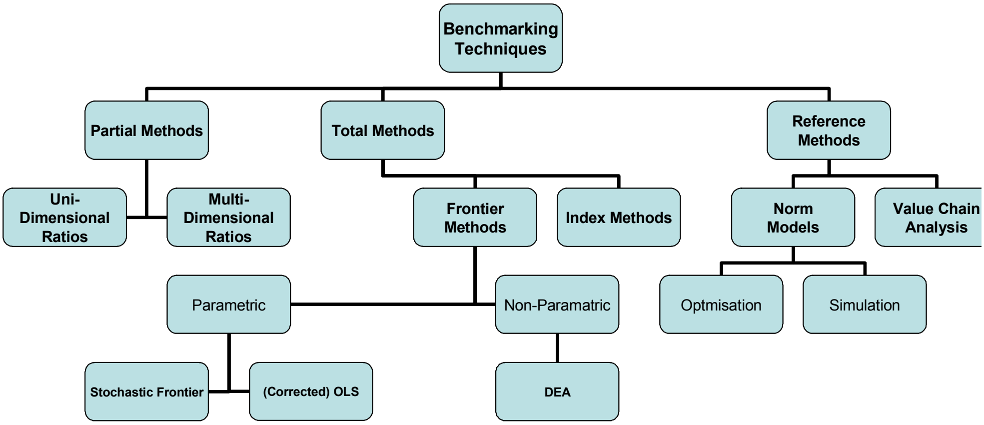


Figure II-1. Overview of benchmarking techniques.

Figure II-1 provides an overview of different benchmarking techniques. The simplest form of benchmarking is to consider uni-dimensional performance. Examples of such indicators include the energy distributed per employee, the costs per unit of energy. By comparing such performance indicators between firms, important indicative information can be obtained in a relative straightforward way. Performance indicators are sometimes also contained in annual reports of firms or used by market analysts because they are easy to calculate and interpret. The main drawback of simple performance indicators is that they fail to account for the relationships between different input and output factors.

This can be overcome by somehow weighting multiple performance indicators into a single one. The problem then is how this weighting should be done, or, more specifically, how to determine the relative importance of each individual indicator. Take the example of combining two performance indicators (the number of staff per customer, and the costs per kWh of electricity distributed) into a single one. Clearly, both these indicators provide valuable information about the performance of the firm. However, it may be that one firm is doing very well at one indicator, but less on another. How can the relative performance of two firms be calculated?

Table II-1. Example of different approaches in weighting multiple performance indicators.

Firm	**A**	**B**
# Staff / Customer (x 1000)	10	6
Costs / kWh (EUR / kWh)	3	8
Weighting	**Overall performance**	
Indicator 1 / Indicator 2	**A**	**B**
25 / 70	4.75	7.50
50 / 50	6.50	7.00
75 / 25	8.25	6.50

As can be seen, the weight attached to each indicator greatly affects the overall performance of the firm. If a 25/75 weighting is used, firm A performs better than firm B as on overall it uses less combined inputs per combined outputs. However, if the weighting is reversed to 75/25, then the situation changes and firm B now performs best. The main limitation of using performance indicators is that the setting of the weights is not unambiguous.

One way of course is to simply fix the weights, for example by choosing a set of weight factors that generally are considered appropriate. This is done when using index methods; these define an index as an aggregation of the ratio of all output quantities (usually weighted by revenue shares) and all input quantities (usually weighted by cost shares). This is called the Total Factor Productivity (TFP). Given the arbitrary character of the weights, the method would not be credible to evaluate relative performance in a single year. Rather, index methods aim to observe changes in productivity over time on the basis of the same choice of inputs, outputs, and weights. This restricts its usefulness as a benchmark technique in the regulatory context as it takes a number of years before the results from such analysis can be used. Furthermore, the heavy data requirements and problem of establishing appropriate weights make the method less practical. However, in the case of yardstick competition, TFP can be used for regulatory purposes. The Dutch application of index methods to calculate the yardstick for price and quality is an example of this approach.

A different approach to determining the optimal input and output weights are frontier methods. As productivity is defined as the ratio of (weighted) outputs over (weighted) inputs, the frontier can be thought of to consist of firms that use the optimal combination of inputs and outputs. Thus, a suitable set of weights can be derived from information about firms operating at the frontier. There are two main techniques to do this. Firstly, non-parametric methods do not impose any functional form on the relationship between inputs and outputs. The most common non-parametric approach is Data Envelopment Analysis (DEA), under which methodology the frontier is made up of linear combinations of the best performing firms in the sample. Secondly, parametric methods impose a functional form on the frontier using estimation for production or cost functions. They require more knowledge about the production or cost functions and about the distribution of errors. However, to test for the validity of the assumptions and to fine-tune the weight assigned to each variable a large number of networks are required.

Parametric frontiers could be estimated by some variant of Ordinary Least Squares (OLS) or by Corrected Ordinary Least Squares (COLS). Under OLS, the frontier is based on the average cost function while COLS tightens the criterion and shifts the frontier towards the best performing firm. Stochastic Frontier Analysis (SFA) attempts to estimate an efficient cost frontier that does incorporate the possibility of measurement error or chance factors in the estimation of the efficient frontier. This method first allows for the adjustment of individual costs for stochastic factors and then calculates efficiency scores in a way similar to COLS. The efficiency scores are usually higher than under the COLS method precisely because the most efficient firm under COLS will be assumed to be subject to some negative stochastic factor affecting its actual costs. While this method incorporates stochastic factors, it still requires the specification of a functional form for the efficient frontier. It further requires the specification of a probability function according to which the stochastic errors are distributed.

The discriminative power of the benchmarking analysis will be lower if the sample size is smaller. Generally, a rule-of-thumb is that the sample size should be larger then three times the total number of input and output factors that are considered. So, for example, when opex is chosen as an input and the number of customers and energy distributed as an output, the minimum size of the sample should be minimal $3 \cdot (1+2)=9$.

If the sample size is smaller, it is less likely that the results from the benchmarking analysis will identify the firms' true efficiency (potential). In that case, another approach can be chosen. Rather than comparing the relative productivity between firms, one could artificially construct an ideal reference for the firm(s) and relate efficiency to their performance relative to this benchmark. This is known as the reference model. This approach is similar to benchmarking with the notable exception that the firm against which is being benchmarked is not an actual firm but an artificial firm, constructed on the basis of certain assumptions.

Annex III. Quality and Density Data

This Annex contains the data from Figure 4-5. Note that SAIDI figures may not be entirely comparable due to differences in definitions or measurement. Data is collected from annual reports and the website of the respective firm.

Firm Name	Country	City / Region	Area size [km2]	Number of connections	SAIDI [min/year]
BEWAG	Germany	Berlin	890	1,800,000	14
ConEd	USA	New York City	1,709	3,100,642	23
Delta Netwerk	Netherlands	Zeeland	3,117	188,350	20
East Midlands	UK	Mideast England	16,000	2,300,000	71
Edelnet	Netherlands	Delft	162	116,197	21
EDON	Netherlands	Northern Holland	9,454	890,480	25
ENBU	Netherlands	Central Holland	1,369	494,865	33
Eneco	Netherlands	Rotterdam area	2,212	1,075,176	53
Energy Australia	Australia	NSW / Sydney	22,275	1,441,000	118
Enet	Netherlands	Eindhoven	87	98,509	69
ESB	Ireland	Dublin	70,000	1,600,000	385
EZK	Netherlands	Kennemerland	10	11,600	56
Frigem	Netherlands	Friesland	84	45,247	7
Integral Energy	Australia	NSW / South of Sydney	24,500	761,000	218
KEPCO	South Korea	Seoul	98,480	15,619,000	19
London Electricity	UK	London	665	2,000,000	42
MEA	Thailand	Bangkok	3,192	2,500,000	98
MEGA	Netherlands	Limburg	2,009	405,119	12
NMHO	Netherlands	Gouda	324	86,938	20
Northern	UK	North England	25,000	3,600,000	91
NUON	Netherlands	Central / Northern Holland	17,456	2,614,059	25

Firm Name	Country	City / Region	Area size [km2]	Number of connections	SAIDI [min/year]
PG&E	USA	San Francisco	181,300	4,500,000	166
PNEM	Netherlands	South Holland	4,645	929,072	8
PowerGrid	Singapore	Singapore	682	1,165,280	4
Seeboard	UK	Kent, Sussex, Surrey	8,200	2,100,000	75
Taipower	Taiwan	Taiwan	32,260	10,746,000	72
Tenaga	Malaysia	Kuala Lumpur	131,598	5,027,128	330
TEPCO	Japan	Tokyo	39,496	26,669,000	3
Western Power	UK	West England	26,000	2,400,000	129

Annex IV. Case Study: Italy

Introduction[1]

The objective of this case study is to evaluate the effectiveness of the Italian quality regulation as applied during the period 2000 till 2003. This is done by studying the effects of regulation on the electricity distribution industry and to identify scope for improvement of the regulation in subsequent regulatory periods. As part of this case study, a series of interviews was conducted with different key-players from Italian electricity distribution firms, consumer associations, electricity industry associations, government agencies, and research institutes. The interviews took place in Italy (Milan, Modena, Rome, Turin, and Vercelli) during May and June 2003.

The case study starts with a description of the regulatory framework in Italy and more specifically, of the quality regulation system. Subsequent sections then move to the results of the evaluation. This starts from the assumption that there is a causal relationship between regulation and the observed improvements in continuity levels. The interaction between regulatory policy and network management is then analysed. More specifically, the extent to which regulation has been accepted by network managers and what factors have contributed to this. In addition, there may have been factors other than regulation that have driven network managers to pursue better continuity levels. As will become clear, liberalisation of the electricity markets has played an important role. The impact of regulation on network management is then considered as well as some potentially undesired effects that may (unintentionally) have been caused by regulation.

Regulatory Approach

Italian Regulatory Framework

In Italy, the energy regulator (Autorità) is responsible for setting tariffs and quality standards. According to is founding law (Law n. 481/1995), the main objectives of the Autorità are to guarantee the promotion of competition and efficiency and to ensure adequate service quality standards in the electricity and gas industries. In order to induce electricity distribution firms to operate more efficiently, a price-cap system was put in place in 2000. This system pushed network operators to reduce costs by an annual four percent during four years. The potential dangers of quality degradation under price-cap regulation were recognised at an early stage. In parallel with the price-cap system, the Italian regulator introduced an incentive scheme for electricity continuity of supply to apply during the period 2000 till 2003. Two initial observations defined the shape the Italian quality regulation system. Firstly, a comparison with other EU countries revealed that there was room to improve continuity of supply levels in Italy.[2] Secondly, within Italy itself, there were high geographical differences. More specifically, continuity of supply levels in the South were substantially lower than in the North. These observations led to the two main objectives of the Italian regulatory system namely (1) to increase continuity of supply levels in Italy and (2) to bridge the gap between North and South.

These objectives were materialised in terms of the prescription of national reference standards, which reflect the regulatory judgement of continuity of supply levels Italian distribution firms should operate at. These national reference standards are measured in terms of the annual average minutes of interruption per customer (SAIDI)[3] and vary geographically as a function of customer density. Geographical differentiation of the standards was done in order to recognise differences in operating conditions to deliver high continuity of supply. Generally, the efforts required to produce high continuity levels are inversely related to the level of customer density. Three density levels were identified: Urban municipalities (more than 50,000 inhabitants), suburban municipalities (between 5,000 and 50,000 inhabitants), and rural areas (less than 5,000 inhabitants). The national reference standards were respectively set at 30, 45, and 60 minutes of SAIDI.[4]

Table IV-1. Level of annual improvement target by territory and regulated band

Historical Continuity of supply Levels (1998-99 average)			Annual target improvement rate
Urban Areas	Suburban Areas	Rural Areas	
up to 30 minutes	up to 45 minutes	Up to 60 minutes	0 percent
From 31 to 60 minutes	from 46 to 90 minutes	from 61 to 120 minutes	5 percent
From 61 to 90 minutes	from 91 to 135 minutes	from 121 to 180 minutes	8 percent
From 91 to 120 minutes	from 136 to 180 minutes	from 181 to 240 minutes	10 percent
From 121 to 150 minutes	from 181 to 270 minutes	from 241 to 360 minutes	13 percent
more than 151 minutes	more than 271 minutes	more than 361 minutes	16 percent

Distribution firms that were not already operating at the national reference level were required to improve annually towards these standards (see table IV-1). The annual improvement targets range from zero to 16 percent with national average of about 10 percent. The targets were set based on the density area and historical levels of continuity of supply: the improvement target was higher for districts with lower quality levels; the improvement target was set at zero for districts where national reference standards were already reached. Furthermore, a direct link between the level of continuity of supply delivered and the allowed income to the firms was introduced. In case the firm did not manage to improve continuity levels according to the improvement target, it pays a penalty. Conversely, if the firm improved better than the target, it received a reward. The level of the penalty or reward is proportional to the difference between the actual performance and continuity target. The indicator was chosen to be the two-year rolling average for SAIDI in order to reduce as much as possible distortions from random effects. In this spirit, an additional symmetrical dead-band of five percent around the target was applied.

Eventually, the incentive scheme for continuity of supply could be constructed for each district based on its historical performance and density level in the following way. Let us assume that T_i is continuity target for the two-year period $(i\text{-}1, i)$, A_i is the actual level of the SAIDI indicator as two-years rolling average for the same period $(i\text{-}1, i)$, and E_i is the total consumption of Medium- and Low Voltage users belonging to the district in the year i, and R^d the relevant national reference standard for the grade of density d, and db is a constant dead band of five percent. Incentive and penalties for the firm in the given district and in the two-year period $(i\text{-}1, i)$ are then defined according to the following formulas:

- The case $A_i < S_i*(1\text{-}db)$ implies that the firm has improved continuity more than required by the target S_i; in this case, the firm gains $G_i =[S_i – max(A_i; R^d)] * E_i/8760 * C >0$ (the positive sign means that G_i is a reward to be received by the firm);
- the case $S_i*(1\text{-}db) < A_i < S_i*(1+db)$ implies that the firm has improved continuity at the required rate and within the boundaries set by the dead band. In this case, the firm does neither gain nor lose anything (“dead band” effect);
- the case $A_i > S_i*(1+db)$ implies that the firm has not improved continuity as required. In this case, the firm is imposed a penalty equal to $G_i =[S_i \text{-}A_i] * E_i/8760 * C <0$ (the negative sign means that G_i is a penalty to be paid by the firm).

In the cases a) and c) respectively, the level of the reward and penalty are determined based on the parameter C, which varies according to the density and the continuity level of the district under consideration (see table IV-2). On average, the parameter C corresponded to approximately 18 Euro per kWh of non-supplied energy, which constitutes a strong incentive to outperform or at least respect the continuity targets. On the other side, given that the parameter C is also used to set the penalty in case of underperformance, it provides firms with an incentive to act according to the continuity targets in order to avoid paying high penalties.

Only a sub-set of the total number of interruptions was initially considered in the regulation. One reason to do so was that it would be unfair for the regulator to punish the network operator for interruptions resulting from events that could not have been anticipated. These so-called Force Majeure events mainly relate to exceptional natural phenomena like floods, earthquakes, etc. Another category of interruptions that are excluded are those caused by third parties (e.g. digging activities or sabotage) or those that originate outside the distribution network i.e. in the higher transmission networks or those related to generation interruptions.

Table IV-2. Values of parameter C by density and regulated indicator bands

High-density territorial districts	Medium-density territorial districts	Low-density territorial districts	Parameter C (ct EUR/minute/kW)
up to 60 minutes	up to 90 minutes	up to 120 minutes	41.3166
from 61 to 120 minutes	from 91 to 180 minutes	from 121 to 240 minutes	30.9874
more than 121 minutes	more than 181 minutes	more than 241 minutes	20.6583

The incentive system is funded through penalties paid by utilities for those districts where the continuity targets are not met, and for the net difference between incentives and penalties, through a Q-factor in the price-cap formula:

$$price\text{-}cap=RPI-X+Q$$

The Q-factor is calculated ex-post, and may assume a negative or positive sign. In the case $Q>0$, it means that as a whole the system improved more than required, and all users are called to contribute. On the contrary, $Q<0$ means the whole system improved less than required (on average, less than 10 percent a year), and all users benefit of a reduction in tariff. In Italy, a uniform tariff applies making it necessary to maintain an equalisation fund to distribute incentives to utilities that are linked to different levels of quality without changing the final tariff district by district.

Quality Measurement

Uniform measurements and data recording systems as well as audit systems are important to assure that the data used in the regulation system are correct and do not lead to erroneous regulatory outcomes. Although the calculation of the continuity levels is a straightforward process, comparison of performance data between firms or countries should always be treated with caution. Experiences in (for example) the UK and the Netherlands show that differences in definitions and measurements can lead to distorted comparisons (Ofgem 2001, DTE 2002b). For example, sometimes the term 'interruption' itself is not uniformly defined. Some only include an interruption in the statistics if it lasts longer than three minutes, while others include it if it lasts longer than one minute. The starting point of an interruption may also not always be defined in the same way since some interruptions are detected automatically, while others are recorded at the time that customers start ringing the firm to report the interruption.

The Autorità devoted substantial attention to assure uniformity in definitions and interruption recording processes. Distribution firms were obliged to set in place adequate recording systems and to follow uniform guidelines. Interruptions were classified between long interruptions (longer than three minutes), short interruptions (between three minutes and one second), and transient interruptions (shorter than one second). During the first regulatory period, transient and short interruptions were not included in the regulation but from 2002, after having allowed distribution firms time to properly implement data recording systems, statistics are collected even on transient and short interruptions. Network operators were required to record interruptions both automatically (using the control and monitoring systems) and manually (for restoring operations). For high and medium voltage lines, there is a requirement for remote detecting and recording of interruptions. A similar requirement has not been made for the low voltage network because of the high costs involved and the relatively small overall effect on the continuity indicators. Firms are furthermore required to submit an annual report on continuity of supply performance to the Autorità. This report includes the continuity of supply indicators for each district, differentiated by cause, origin, and type of interruption. These indicators and their underlying data are audited by the Autorità in order to make sure measurements are accurate and comply with measurement requirements.

Evaluation of the Italian Quality Regulation

Evaluation Methodology

When analysing the statistics on continuity of supply levels, one may observe a significant improvement during the period 1999-2002. In these years, the national average level of system average interruption duration index (SAIDI) improved from 228 to 130 minutes i.e. improved by 43 percent in three years (see fig.1). The frequency of interruptions – as measured through SAIFI - also improved with 30 percent from 4.2 to 2.9 interruptions per customer. These figures include all interruptions, also those currently not considered in the regulation i.e. external and Force Majeure events. The largest improvements have taken place in the Southern regions. As a result, the gap between Southern and Northern regions has decreased as much as 60 percent. Furthermore, for each of the three density levels, the gaps between the 90° percentile and 10° percentile of the distribution of SAIDI have been remarkably reduced. For instance, for high-density districts, this gap reduced from 120 to 50. The number of districts where continuity of supply levels were already at or above the national standard increased from 19 in high density, 15 in medium density and 3 in low density to respectively 42, 19 and 4.

The statistics on continuity performance in the last few years suggests that levels have developed well in line with the predefined regulatory objectives. It is plausible that there is a causal relationship between regulation and these outcomes. However, this does not reveal whether it has been regulation alone that drove the improvements and the channels through which these have been achieved. More insight into these issues can be very valuable to the regulator to further improve the regulatory system. These issues are explored in the next sections, which report on the results from the field study. As will become clear, network managers have generally acknowledged the need for quality regulation and the objectives as formulated by the regulator.

This section deals with the underlying reasons for this and the role of market liberalisation in this regard. Further sections look at the concrete measures taken by distribution firms to improve quality. Also, some of the potential adverse effects that need to be dealt with in future regulatory periods are analysed.

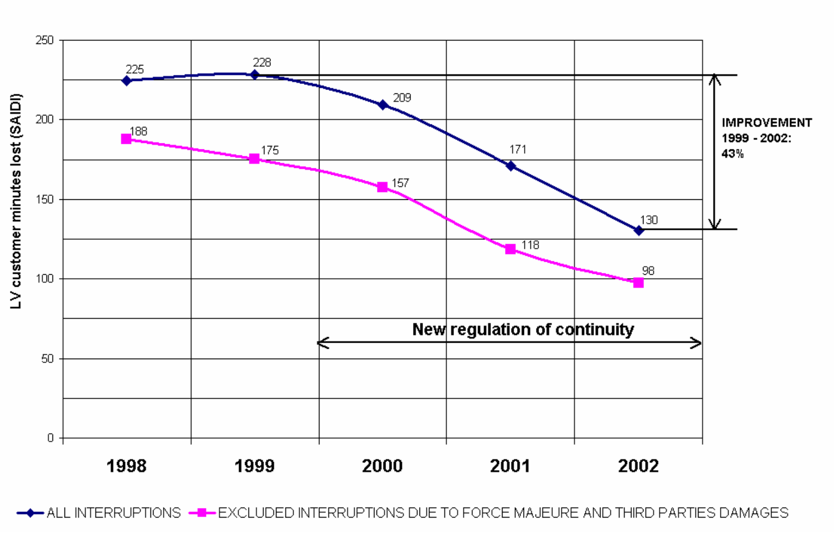


Figure IV-1. Effects of continuity regulation in Italy. Developments in SAIDI levels.

Key Success Factors

The need for higher continuity of supply levels in Italy, in particular given the historical low performance relative to other European countries, had been widely recognised by network managers. On the other hand, the system of uniform tariffs in Italy strongly contrasted with the large regional quality deviations within Italy. In that light, the regulatory policy of higher and more uniform quality levels was well received and supported by distribution firms. Indeed, there was a general awareness that continuity levels should improve well before regulation came into place. However, there was no clear idea about how much this improvement should be while the absence of formal requirements to measure and improve quality made it difficult for firm managers to set in place improvement programs with clearly defined targets and suitable policies to meet them. The introduction of formal quality regulation materialised the general idea of quality improvement. It provided a uniform measurement system and tangible targets with associated financial incentives. Thus, although obligatory, firms' managers were well receptive of quality regulation as it was in line with their own internal policies. Regulation did not only make matters more concrete for firms, generally it also was perceived as fair as it recognised the differences in historical and operating conditions through differentiation of improvement targets by customer density and historical performance levels. Furthermore, applying a symmetric incentive scheme, i.e. not only penalties for under-performance but also rewards for over-performance - strengthened the acceptance of the system.

As will be shown later, a large part of the continuity improvements have been achieved through increased capital investments. Investments in electricity distribution continuity generally do not increase sales and therefore are not profitable.[5] Strictly speaking then, distribution firms do not have a natural incentive to increase quality unless they are provided with a financial incentive to make these quality investments.[6] Under the Italian regulation scheme, this is done by adjusting the regulatory asset base (RAB) to reflect quality related investments. A higher RAB is associated with higher allowances for depreciation and returns. From an investor point of view, quality investments are then favourable as they are not only recovered, but also a return can be earned on the invested capital.

Quality Reputation Spill-Over Effect

The interviews revealed a general agreement that regulation has been an important source for improving continuity levels. However, equal importance was also attached to the increased consumer demand for higher quality in combination with the further liberalisation of previously captive supply markets. In addition to regulation, this has been an important driver for distribution firms to supply better levels of continuity of supply. At first glance, this may seem counterintuitive as economic theory predicts that monopoly providers are not likely to act in line with the interest of consumers. However, the explanation for this effect can be found in the spill-over of quality reputation from the monopoly towards competitive business areas. When a monopoly firm is also active in complementary but competitive markets, the incentive to reduce investments in quality under price-cap regulation might be tempered (Weisman, 2002). The explanation for this is that a reputation of poor quality in provision of monopoly services spills over to the businesses in the competitive markets. Rather than decreasing quality in the monopoly markets, the firm is therefore pressured to bring quality levels in line with customer demands. Failing to do so can lead these customers to choose another supplier in the other complementary competitive markets and in turn reduce the firm's overall profitability.

The Italian case is a good example of this spill-over effect. Electricity distribution firms in Italy are usually part of a larger holding firm which also operates in competitive markets such as electricity generation and supply as well as in other (partially liberalised) sectors such as gas, water and telecom. Performing well in the monopoly network business improves the image or branding value in the competitive businesses and strengthens the overall competitive position. Our research indicated that this has been an important reason for network managers to comply with the demands for higher quality – both by customers and by the Autorità. Another related effect is that not meeting the regulatory targets does not only result in a financial penalty but also creates substantial damage to the firm's image. Indeed, as the interviews revealed, this image loss was sometimes found to be more important than the financial penalty.

We may argue that in Italy the spill-over effect has played an important role in the drive towards better continuity levels. It is, however, difficult to say if the effect would have been as strong in the absence of regulation. As showed before, regulation created the framework – measurement systems and robust targets – for distribution firms in order to set in place proper management policies for continuity improvement. It is unclear whether the spill-over effect in isolation would have produced the same results one may observe today or that it has rather amplified the incentives for distribution firms to improve continuity. Either way, an important observation remains that setting quality objectives in line with market demands can help to strengthen the effectiveness of the regulation.

Impact on Network Management

Quality Investments

Network managers have taken different actions to comply with the regulatory quality requirements. Our evaluation revealed that most of the efforts in Italy to improve continuity of supply have been capital driven. Investment levels have substantially increased since the beginning of the quality regulation with increases up to 50 percent. In particular, in the Southern parts of Italy where continuity supply levels have been historically lower, investments in quality improvement have been substantial. The majority of investments have aimed at reducing the interruption duration; frequency of interruptions has received less priority. The likely reason for this is that it is relatively easier for a distribution firm to reduce the average duration rather than frequency of interruptions. Measures such as automation are relatively easy to implement but can significantly reduce interruption duration making the impact on the SAIDI indicator more profound. The main area where such investments have taken place is the installation of secondary telecontrol equipment along the Medium Voltage lines in order to perform switching actions from a central location (telecontrol at the start of each High- and Medium Voltage line is a regulatory requirement). Through secondary telecontrol, interruption restoration time is reduced as travel time is eliminated. In addition to the improvement in continuity of supply, automation also reduces the labour costs. In general, network managers have aimed to automate 30-40 percent of all Medium to Low voltage transformer stations making this an important effort to improve continuity levels.

Another important measure taken in response to regulation has been the installation of Petersen coils, which limit the short circuit currents resulting from faults in the distribution network. The number of long interruptions can be reduced as faults become transient rather than a sustained interruption. Typically, firms aimed at installing Petersen coils in all High Voltage substations. The costs of Petersen coils are moderate making it a cheap but yet effective method to improve continuity of supply given the actual conditions of Italian distribution networks. Undergrounding has been another measure taken by firms - underground cables are less sensitive for external effects (e.g. weather) leading to higher network robustness and less interruptions. Especially in areas with more severe weather conditions, undergrounding can substantially improve continuity of supply. Undergrounding will mainly bring down the frequency of interruptions but at the same time can have a less favourable impact on the duration. In the case of radial networks – when re-routing is not possible – this will increase the repair time and thus the duration of the interruption. Due to its high costs and limited effectiveness, undergrounding has only been applied in some instances.

Finally, another area where improvements have been introduced is that of interruption management. Here, the changes have been mostly in the better organisation of interruption crews. A noticeable trend is a more decentralised locating of interruption crews in order to improve dispatching of personnel to the faulted areas. In some cases, the number of staff has been increased, trained more intensively and provided with equipment to be able to navigate more efficiently (e.g. GPS systems). For larger firms, interruption crews operating in one area are now expected also to operate in different (neighbouring) areas in order to exploit economies of scale.

Data Measurement

The availability of uniform data measurement systems has been extremely important to assure the credibility of the calculation of the regulatory penalties and rewards. Previously, distribution firms often kept records of interruptions for purposes such as maintenance and investment decisions. Recording was, however, limited and measurement systems and definitions tended to vary between firms and even between departments of the same firm. The obligation to properly record interruptions according to the Autorità's guidelines has significantly improved this situation. Substantial human and capital resources have been invested in interruption recording systems. In most cases, one or more persons have been trained and assigned the task of data recording and reporting. In addition, specific software has been developed or purchased to automate the data recording process. Now, standardisation and improvement of the measurement processes in combination with the auditing process have resulted in high quality continuity of supply statistics.

However, some problems, however, remain with the measurement of interruptions. In particular, the definition of Force Majeure remains a topic of discussion between firms and the regulator. During the first regulation period, the Autorità only considered an interruption to be the result of Force Majeure in case of events that have been officially classified as such by the national government. Firms, however, thought this as a too strict criterion and considered less severe events as Force Majeure. These conflicting views have led to substantial debates between the regulator and firms. Another remaining problem is the auditing of data by the Autorità. In order to make sure that performance is measured correctly, data audits are performed, and quality of the measurement system is evaluated through accuracy and precision indexes. These indexes are, however, quite sensitive in case of few interruptions due to the normalisation by the number of interruptions. This is particularly true for small firms that supply high continuity levels.

Customers Perception

Information on continuity of supply statistics have been made publicly accessible by the Autorità, increasing the transparency of firms' performance and supporting customers or customer representation groups in discussions with network operators. Access to information about the level of firm's performance and the ability to compare this with other firms (either nationally or internationally) helps customers to judge whether their received performance is adequate and if not, put directly or indirectly apply pressure on the firm to improve. The informational position of consumers in terms of being informed about performance levels of the firms was rather limited before regulation. Monitoring quality levels and making the results transparent has helped consumers to reduce this informational gap. It remains questionable whether individual consumers are actually aware of the benefits of regulation. Regulation may have been a success at the macro level but there is no evidence that it has been perceived as such at the micro level. Consumers are not likely to be able to detect the improvements that have taken place – in particular when continuity levels are already high, the marginal improvement is hard to detect. One way to make the benefits more tangible to consumers is to introduce direct financial compensation in case of interruptions – a measure that is to be implemented in future regulations.

Potentially Undesired Effects

From the evaluation, it has become clear that network managers have invested significantly in order to improve continuity levels in line with the regulatory targets. Mainly, improvements have targeted at increasing investment levels and reducing the duration of interruptions. The frequency of interruptions has received a much lower priority, as it is more difficult to improve here. It is, however, important that both duration and frequency receive sufficient attention. Different empirical studies show that the cost of an interruption exists of two elements: A fixed component, which is independent of the duration of the interruption, and a variable component that increases with the interruption duration.[7] The fixed component can be high, in particular for industrial consumers producing high-sensitive electronic components. This means that the frequency of interruptions has substantial impact on the total interruption costs and should not be ignored. It may be true that initially the highest gains can be reached through duration reduction, but in a later stage, the importance of frequency will become more pertaining.

An issue related to the previous point is the distinction between transient, short, and long interruptions. Interruptions with duration shorter than three minutes currently do not count in the statistics. The reliability improvement measures cause relatively more interruptions now to be classified as short or transient. Consequently, these interruptions no longer influence the financial incentive. For example, by applying telecontrol, interruptions that previously would have lasted a few hours, would now last only a few seconds. Not including these anymore in the regulation can lead to strategic allocations in terms of long, short and transient. In addition, there may be a perverse incentive to ignore the frequency of interruptions as long as these remain within the three-minute threshold and thus do not affect the outcome of the regulation. This may result in a situation of shorter interruptions, but occurring at a higher frequency.

Conclusions

This evaluation study confirms that the quality regulation of the Autorità has largely achieved its objectives and has exposed the ways through which this has been achieved. On the other hand, the evaluation also identified some potential problems that are due to be addressed in the next regulatory periods. These are the inclusion of explicit interruption frequency indicators and targets in the incentive scheme, introduction of direct compensation payment to consumers, redefinition of the indexes used to evaluate correctness of the data, better definitions for Force Majeure events, and the development of ideas for future regulation once firms already operate at the prescribed national reference standards. Evaluating the Italian quality regulation system also contributes to a better understanding of the effects of quality regulation of electricity distribution networks. As discussed in the introduction, there is a general awareness of the need of quality regulation under stricter forms of price controls. The theoretical arguments for this need and the problems faced in developing effective quality regulation systems are well known. However, relatively little experience has yet been gained with practical application of quality regulation. It is therefore important to point out some lessons one may learn from the Italian experience.

Firstly, designing a regulatory system that was inherently compatible with the internal objectives of the distribution firms strengthened the acceptance and effectiveness of the system. By defining national references standards and improvement paths, distribution firms were provided with a clear framework of where quality levels were eventually to arrive at. At the same time, it was recognised that in order to achieve these targets, it was required to induce firms to invest in improving quality and to remunerate these investments by including a quality component into the price-cap formula. Proper handling of operational issues has also contributed to the success of the regulation. By investing in a robust data recording system firms were convinced that the data used in order to set penalties and rewards was indeed measured appropriately within and between the network operators. This reduced discussions on data correctness and both the regulator and firms were able to focus on the real regulatory issues of improving and harmonising quality throughout Italy.

Secondly, the spill-over effects of quality from the monopoly to competitive markets have contributed to the effectiveness of the quality regulation. Distribution firms have been sensitive for consumer demand for higher quality as this also affects their profitability at other competitive business areas. Thus, in the Italian case, it may be fair to conclude that the absence of a requirement to *fully* separate network business from the competitive ones, has positively affected quality performance. Furthermore, the benefits of the spill over effect may be best exploited if there is a clear regulatory framework in place. This is not to say that unbundling in general is undesirable. Unbundling requirements are important to accommodate competition. However, the findings from our study also show that in a country where proprietary unbundling has not been imposed, some positive spill-over effects have been evidenced which have provided incentives to supply adequate quality levels.

Finally, the evaluation research confirms that quality regulation is still an innovative part of regulation and requires periodic evaluation and verification of its outcomes - both desired and undesired. Furthermore, quality incentive schemes are inherently *country-specific*: they should be designed contingent on the actual industry and regulatory framework of the country. A "superior" quality incentive scheme, fitting each possible contingency, probably does not exist. Nonetheless, evaluation of national schemes – as well as international benchmarking exercises comparing actual quality levels and standards[8] – can help to build an international "policy community" around quality regulation in order to compare different experiences, to identify common issues and to let practicable solutions to common issues emerge.

The main objectives of the quality regulation system discussed in this case study were to improve continuity of supply levels and to even out geographical differences. As has observed, these objectives have generally been achieved. The question then arises whether these positive outcomes are also sustainable in the longer-term and if not, which measures should be taken to make this happen. This provides some directions for further research.

Firstly, regulatory objectives may need to be modified over time. It is generally accepted that higher quality involves more costs and hence higher prices. Allowing investments to be included in the RAB promotes the success of the quality regulation as firms have a strong incentive to engage into the necessary network investments to improve continuity levels. At some point however, the costs of extra quality will not outweigh the consumers' willingness to pay for this extra quality.[9] Then, further quality improvement may not be desired and rather, one should move to an incentive structure that aims at optimal rather than higher levels of quality. This requires an adaptive regulatory system i.e. a system where regulatory objectives and approaches change in line with the changing quality requirements.

Secondly, the possibility of a too long lag between quality decisions and quality outcomes should be taken into account. Given the capital intensity of the electricity distribution business, there may be a long period before (some) cost decisions materialise in terms of a change in quality levels. As has been shown in this case study, quality levels may be substantially improved by making the right investment choices. However, in the context of price-cap regulation, network managers may have an incentive to adversely save on quality costs and the effect of some of such actions may only become visible after some time. If this period is too long, then managers may not be sensitive for a quality incentive scheme that focused on one regulatory period only. That is, their appointment term may well be shorter than the time it takes for their decisions to materialise in terms of lower quality and subsequently a penalty. A further study of the dynamic relation between costs and quality may reveal information about the length of the time lag and can help in developing regulatory systems that lead to sustainable quality outcomes also in the longer term.

Notes

[1] This case study is based on Ajodhia et al. (2006).

[2] See for example CEER (2001)

[3] SAIDI stands for System Average Interruption Duration Index.

[4] Two-years average, net of interruptions caused by Force Majeure or by third parties' damages or customers' faults and of interruptions originated in the EHV transmission networks (>35 kV)

[5] Increments in continuity may increase sales as a result of higher output but this effect is small and can safely be ignored if variations are not large (Munasinghe, 1979 p.36).

[6] In reality there may be other incentives to supply high quality even if it would not be profitable to do so. These include spill-over effects which will be discussed further.

[7] See Ajodhia et al. (2002) for an overview of electricity interruption cost studies.

[8] See for example CEER (2001), CEER (2003) and CIGRÉ (2004) for international comparisons of quality regulation of electricity distribution.

[9] For electricity networks, it is generally assumed that, at the margin, costs to supply higher quality increase but consumers' willingness to pay decreases at higher quality levels. This implies that an optimal quality level exists. Here, costs and willingness to pay for quality are in balance. See Munasinghe (1979) for a discussion of optimal quality concepts.

Annex V. Case Study: Norway

Background

Institutional Setting

Deregulation and market competition were introduced in Norway with the Energy Act of June 1990. The necessary organisational structure was completed in May 1992 and since that time Norway has had an open competitive electricity market. The common Norwegian-Swedish Exchange, the first electricity market completely open to trade across national borders, has been in operation since January 1996. A joint electricity market for the four countries of Norway, Sweden, Denmark, and Finland was recommended in August 1996 and covers all Nordic countries from July 1999.

The transmission and distribution of electricity in Norway is carried out by a large number of firms. Transmission is carried out at three different levels: via the main network, regional networks, and distribution networks. Statnett SF, a state-owned enterprise, owns by far the largest part of the main network and is responsible for tariffs, system operations, and the development of the main network system. The main transmission network includes 400 kV, 300 kV and 132 kV. Statnett owns about 77 percent of the Transmission Network and leases the remaining 23 percent. Some 40 other distribution firms (regional firms and producers) each own small sections of the main network. Statnett SF has a leasing agreement with these 40 firms, and the lease costs historically are passed on to consumers. Between 50 and 60 firms are involved in the transmission of electricity at the regional level. These firms are often vertically integrated in the sense that they also produce and sell electricity. They are also often involved in the distribution of electricity at the local level. The regional networks are often owned by local and/or regional authorities. Electricity is distributed locally by around 200 firms, often owned by the local municipalities. These firms vary greatly in size and other characteristics. The average distribution firm has around 5000 customers. Some of the distribution firms feature local production. The majority of the distribution firms are also engaged in the sale of electricity, mostly to local customers.

Regulatory Framework

The Norwegian Water Resources and Energy Directorate (NVE) is the power industry regulator in Norway. The NVE is a directorate under the Ministry of Petroleum and Energy, with responsibility for managing the country's water and non-fossil energy resources. NVE's mandate is to ensure integrated and environmentally friendly management of the country's watercourses, to promote efficient energy markets and cost-effective energy systems and to work to achieve a more efficient use of energy. NVE is responsible for reducing damage caused by floods and erosion along rivers. NVE has important duties in the national preparedness against floods and other water-related disasters. NVE also has the overall responsibility for maintaining national power supplies.

Network regulation in Norway was based on rate-of-return regulation during the years 1992-1996. However, the deficiencies resulting from this approach were soon recognised. The main issues were inefficiencies caused by the guaranteed cost recovery and the weak incentives for productivity improvement. This became the major reason to replace the existing price control framework by a new incentive-based scheme in 1997. The current Norwegian regulatory system is an ex ante regulation method based on incentive regulation with the help of income frames. Through incentives, NVE strives to encourage network owners to reduce costs and improve their efficiency. Under the new system, network owners are no longer guaranteed full cost recovery. By establishing a system where each network owner is allowed to receive pre-determined maximal revenue, the profit will in principle be the difference between allowed revenue and actual costs. Allowed revenue requirements should cover the networks' total costs: operation and maintenance, capital costs in the form of depreciation and return on capital invested, network losses and profit tax.

The new regulatory model was based on revenue-cap regulation, supplemented by benchmarking and profit sharing mechanisms. Initial revenue-caps were determined on the basis of the distribution firms' accounts from 1994 and 1995. In 2002, NVE reset the price control for the networks keeping the general logic of the revenue-cap from the first regulatory period (1997-2001), however, adjusting some of the components and seeking improvement of the regulatory cap properties. Additionally, NVE explicitly addressed quality of supply in an integrated price control framework. A link was introduced between the allowed revenues and the level of performance – the so-called CENS arrangement.

Price Regulation[1]

Norwegian revenue-cap regulation in electricity distribution contains elements of different regulatory mechanisms. It consists of cost-plus regulation with a time lag, and benchmarking plays a crucial role in determining efficiency requirements. The revenue-cap is determined by:

- The revenue-cap of the preceding year, or primarily the costs in the first year of the regulation period, plus a standard return on capital for the same year;
- An expected efficiency improvement parameter, benchmarking-based; and
- An annual correction factor intended to provide additional revenue as a function of pre-specified revenue drivers.

The fact that the revenue-cap is affected by these factors makes it possible for the grid owner to influence its return, not only by decreasing costs but also by operating and maintaining the grid in such a way to benefit the revenue-cap. The initial revenue values consist of the average operating and maintenance costs, depreciation, a rate-of-return on invested capital (book value plus one percent to allow for working capital), and average grid losses. The market price for power is used to assess the value of grid losses.

In the first revenue-cap period (1997-2001), the initial revenue-caps were determined on the basis of the distribution firms' accounts from 1994 and 1995, according to the following formula:

$$IT_e = DV + AVS + AVK + NT \tag{V-1}$$

Where IT_e is the initial revenue-cap determined by operating and maintenance costs (DV), depreciation (AVS), return on invested capital (AVK) and costs associated with energy losses (NT). The dynamic time adjustment of the allowed revenue for the grids was based on the following formula:

$$IT_t = ((IT_{t-1} - NT_{t-1}) \cdot \left(\frac{KPI_t}{KPI_{t-1}} \right) + NT_{MWh} \cdot P_t) \cdot (1 - EFK) \cdot \left(1 + \frac{\Delta LE_t}{2} \right) \tag{V-2}$$

Where IT is the starting revenue or costs for year *t*, KPI is the (variation in the) consumer price index, EFK is the efficiency requirement calculated at the beginning of the regulatory period by means of Data Envelopment Analysis (DEA), and ΔLE is the percentage increase in transported energy on a year-on-year basis.

The initial revenue is annually adjusted for inflation, required efficiency increase and by the term $(1+1/2.\Delta\ LE_{a,n})$. The latter is designed to provide additional revenues to the regulated distribution firms that would contribute to the additional opex and capex incurred by the firms because of the increasing volume of transported energy. The anticipated efficiency improvement includes:

- An individual efficiency increase component – measured via DEA on inter-firm comparison of totex; and
- A general efficiency increase component – imposed exogenously by NVE and reflecting the general technological improvement in the industry.

The current revenue-cap scheme (2002-2006), with initial costs values taken from 1996-1999, is slightly different from the above description. Amongst other things, the last term in the formula was removed and replaced by a new term (*Just*) that adjusts for new investment where the increase of transported energy was supplemented by a second driver, namely the relative increase in the number of buildings in each distribution area.

$$IT_t = ((DV + AVS + AVK \cdot r_t) \cdot (\frac{KPI_t}{KPI_{2000}}) + NT_{MWH} \cdot P_t) \cdot (1 - EFK)) + Just_t \tag{V-3}$$

The Norwegian approach relies on the typical incentives to raise the efficiency in the regulatory period by keeping the interim efficiency gains. The regulatory revenue reset is based on the actual costs including checks for deviations between the prescribed revenue path and the actual firms' performance. In addition, the regulated grid providers are exposed to a repetitive benchmarking that is aimed to eliminate the inefficiencies. In the current regulation period (ending 2006), the grid owner cannot influence the revenue-cap directly. The revenue path is decoupled by the actual costs via the application of the regulatory formula.

If one assumed that last year's totex level is the only basis for determining the revenue-cap in the current period, one would impose no cost-reducing incentives on the regulatees. What NVE does instead is to benchmark the opening level of total costs via DEA, and to impose an efficient totex level objective upon all regulated firms throughout the duration of the regulatory period.

The efficiency requirement term contained in the formula *(EFK)* is based on a totex DEA comparison of all distribution firms (there are more than 150 in Norway, albeit decreasing in numbers). As regards asset values, both book asset values and replacement asset values are considered part of the total cost DEA runs, which are then computed twice. Firms are given the 'benefit of the doubt' in that the most favourable DEA scores are used for their revenue requirement calculations after comparing the efficiency score series from the two DEA runs. The relationship between the efficiency requirement and the efficiency measurements based on DEA is sweetened in such a way that the individual requirement will never exceed three percent annually for any distribution firm reporting a DEA cost efficiency score of 70 percent or lower ('efficiency flooring'). Formally, the efficiency requirement target *EFK* is given by the formula:

$$EFK = 1 - (1 - (1 - KE) \cdot 0.3824)^{1/4} \tag{V-4}$$

Where *KE* is the totex efficiency level for any given firm calculated by DEA, which is generously floored at 0.70 for all distribution firms with a reported raw DEA cost efficiency score of less than 70 percent. For a grid owner with regulatory cost efficiency in the floored 70-100 percent interval then, the formula will mean that 38.24 percent of the individual inefficiency in the distribution grid must be recovered over the regulatory time span of four years. Any residual inefficiency will be carried forward to the following regulatory period.

There is an adjustment term in the revenue-cap formula that exceeds the compensation implicit in the incentive to 'beat' the regulatory benchmarking-based cost target. On the assumption that new investment may be caused by an objective need resulting from changes of certain costs drivers as well as that it must be achieved for safety and system security reasons, the investment revenue adjustment element in the revenue-cap formula is intended to give the firms certain compensation for expansion investments in the grid. However, it is important that the adjustment term does not favour unnecessary and/or gold-plated new investments. In other words, the adjustment term should not provide firms incentives to influence their own revenues through uneconomic actions.

New investments involve capital costs such as depreciation and return on invested capital. The majority of such costs are already taken into account by updating the cost base for the revenue-cap periodically. Cost recovery is, however, delayed in time because the updates do not occur continuously. This entails that the net present value of the implied revenues is lower than what would be necessary to cover new capital costs incurred today. The purpose of the investment adjustment term in the revenue-cap is to provide continuity in terms of investment recovery. In addition, new investments may have an impact on operation and maintenance costs, grid losses, and undelivered energy. Such (arguably positive) changes will not result in changed revenue-caps for the firm until the next regulatory review, and must therefore be assessed when determining the level of the ongoing adjustment term. The share of capital costs associated with a new investment that the adjustment term is supposed to cover, depends on:

- The real timing of investments in relation to the four-year update timetable for totex;
- the life time of the investments;
- future inflation;
- future efficiency requirements; and
- the return on capital (discount factor) for the grid owners.

NVE has calculated that between 64.8 percent and 94.5 percent of capital costs are already covered through the four-year totex revenue-cap. The adjustment term shall therefore compensate for between 5.5 percent (100 percent minus 94.5 percent) and 35.2 percent (100 percent minus 64.8 percent) of the capital costs associated with the new investments.

Quality Regulation

Starting on the first of January 2001, NVE introduced a direct link between the allowed revenue and the reliability performance of distribution firms. This quality regulation system is known as the CENS system.[2] The primary objective of the CENS system is to provide distribution firms with incentives to plan, operate and maintain their networks in a socio-economic optimal way and thereby provide a socio-economic optimal level of reliability (Trengereid 2003). The CENS system applies to faults in the network of one kV and higher and which lead to interruptions of longer than three minutes. Reliability performance data is collected through a fault and interruption reporting system developed previously by the industry. The system, named FASIT, has been used in Norway since 1995. Distribution firms currently need to comply with the FASIT specifications when reporting to NVE.

The CENS system can be classified as a continuous quality incentive scheme with both penalties and rewards. Each firm is provided with an individual quality target. This target is defined in terms of an annual amount of cost energy non-supplied. The cost of energy non-supplied follows from the physical amount of energy non-supplied (in kWh) and the specific interruption costs incurred by consumers per kWh of non-supplied energy (EUR/kWh). In determining the specific costs of non-supplied energy, NVE distinguishes between a number of consumer classes as well as between notified and non-notified interruptions. Recently, NVE conducted a new consumer survey to estimate consumer interruption costs. The results of this new survey are shown in Table V-1.

Table V-1. Specific interruption costs currently used in the CENS system. Values are in EUR/kWh. Source: Trengereid (2003).

	Notified	Non-notified
Industrial	5.9	8.5
Trade and service	8.7	12.7
Agricultural	1.9	1.3
Residential	0.9	1.0
Public service	1.3	1.7
Wood processing	1.4	1.7

The difference between actual and targeted performance is the financial incentive the firm is provided with. Essentially, the quality target is an annual amount of interruption costs that all consumers connected to a certain distribution network are allowed to incur. In case that actual quality is the same as targeted quality (in monetary terms) then the incentive is zero. If actual interruption costs exceed the target, then the firm's revenue is negatively adjusted by a corresponding amount. Conversely, if consumers are provided a higher quality level, then the firm earns a reward.

The general idea of the CENS system is that a firm will make an optimal trade-off between its costs and its quality performance. At the one hand, increasing quality will lead to lower interruption costs and hence a reward. On the other hand, this will also lead to higher costs for the firm itself, as it will increase its spending on maintenance, investments, etc. The firm can maximise its total profits by seeking a balance between quality spending and the quality incentive. This balance is driven by the marginal incentive, which is given by the specific interruption costs. It can then be shown that the firm will opt for a quality level that is optimal i.e. reflects a socially optimal trade-off between costs and quality (see also chapter four).

The quality target differs per firm; NVE used two sources of information for establishing the quality target. Firstly, a regression analysis was performed to calculate the expected level of energy non-supplied per firm taking into account factors such as network type, number of transformers, and some climate and geographical factors. Secondly, NVE considered the historical performance levels of distribution firms. The quality target for each firm was set equal to the average of the predicted and actual quality level – as measured in terms of kWh non-supplied energy.

Quality levels normally can change stochastically from year to year. Hence, the costs of non-supplied energy will also vary and this may lead to undesired fluctuations in the level of penalty or reward. To maintain income for the firm at a stable level i.e. to dampen any short-term financial effects of the CENS system, penalties and rewards are not directly translated into revenue adjustments. Instead, the penalties and rewards are cumulated and treated as either a receivable or debt in the firm's financial accounts. The firm has the possibility to change its tariffs within the constraints of the cumulated quality incentives and only after consent of the regulator.

Notes

[1] This section is based on Ajodhia et al. (2005).

[2] CENS stands for Cost of Energy Not Supplied. In Norway, the system is also known as KILE, which stands for Kvalitetsjusterte Inntektsrammer ved ikke Levert Energi.

Annex VI. Case Study: The Netherlands

Background

Institutional Setting

In 1991, the European Communities adopted the European Energy Charter, which promoted competitive markets in the energy sectors. Around the same time, discussions about liberalising the European electricity markets began, which led to Directive 96/92/EC. The Netherlands implemented the directive in the Electricity Act of 1998. This act deregulated the generation of electricity as well as the supply to end-users. The Dutch electricity market is opened in three tiers, the last of which – affecting small consumers – has recently been implemented in the middle of 2004. The Electricity Act requires that management of the networks, which remain monopolistic, should be legally unbundled from competitive business areas, such as generation and retail. This requirement caused a fundamental reorganisation of the previously vertically integrated sector, and resulted in the establishment of two tiers of network firms: The national transmission system operator (TenneT) and about fifteen regional distribution firms. The distribution networks are owned and managed by the successor firms of the former integrated electricity firms. Unbundling of the distribution networks has taken place at the management level. This means that the network and commercial functions can remain within the same holding firm, but with strict divisions ("Chinese Walls") between them.

The Dutch Office for Energy Regulation (DTE) is responsible for implementing the Electricity Act of 1998. As the electricity market is fully liberalised, only the remaining network monopolies are subject to direct regulatory supervision. DTE's main legal rights and duties are contained in the Electricity Act. Within this framework, DTE is specifically responsible for the following tasks:

- Issuing supply licenses for the supply of electricity and gas to captive customers;
- issuing an exemption from the obligation to appoint an electricity grid manager;

- determining the tariff structures and conditions for the transmission of electricity;
- determining guidelines for tariffs and conditions with regard to access to gas transmission pipelines and gas storage installations and, if necessary, issuing binding instructions;
- determining connection, transmission and supply tariffs for electricity and gas, including the discount (price-cap) aimed at promoting the efficient operation of the electricity grid and gas network managers;
- once every two years, assessing whether the electricity grid managers and the gas network managers have met the need for transmission capacity adequately and efficiently on the basis of estimates of the need for transmission capacity submitted by the electricity grid managers and the gas network managers;
- once every two years, assessing whether the license holders are adequately and efficiently able to meet the captive consumers' need for electricity on the basis of estimates of the total requirement for the supply of electricity to captive consumers submitted by the license holders;
- advising the Minister of Economic Affairs on applications for approval of the appointment of electricity grid managers and gas network managers.

Before liberalisation, electricity tariffs in the Netherlands were set by a system that closely resembles cost-plus. Under this system, tariffs were primarily based on observed costs, plus a reasonable rate-of-return. Although much less legalistic and explicit than in the US, the generally observed weaknesses of the traditional rate-of-return regulatory approach also applied to the Netherlands. The 1998 Electricity Act introduced a completely new approach towards price regulation. Currently, tariffs for distribution network use are regulated on the basis of a price-cap system. Tariff levels are annually adjusted by CPI-X, in which CPI is the consumer price index, and the so-called X-factor is the regulator's estimate of future efficiency improvements.

A special feature of the Dutch network regulation system is that both operational and capital costs are considered in the efficiency analysis. The motivation for including capital costs is to enable the regulated firms to make a trade-off between operational and capital expenditures. By also placing capital expenditures under the price-cap, and given that capital costs are a significant portion of total cost, the incentive to improve efficiency not only applies to the short-term but also to the long-term (DTE 2002a).

For the first regulatory period, from 2001 to 2003, DTE's price regulation can be characterised as an ex ante and benchmarking approach. Tariffs are set beforehand for a number of years on the basis of a regulatory judgment of future productivity improvements as disclosed through benchmarking. For both electricity and gas network monopolies, price-cap systems have been used as a basis for setting tariffs. The strategy for setting the X-factors was to drive firms towards similar efficiency levels. This was done to create a level playing field, so that yardstick competition could be introduced in the second regulation period which lasts from 2004 to 2007 (DTE 2002b).

PQRS: Price Quality Regulation System

The yardstick competition system that DTE introduced starting 2004 is also known as PQRS: Price Quality Regulation System. An important feature of the new system is the integrated approach towards price and quality (with reliability as the most important quality feature). The objectives of the new system – which takes effect in 2004 - is to stimulate firms to operate cost-efficient, and at the same time, make a socially optimal cost/reliability trade-off. The regulation system proposed by DTE for the second regulation period is based on the concept of yardstick competition. In competitive markets, profitability of an individual firm not only depends on its own performance, but rather on its performance relative to that of its competitors. The dynamics of competition are simulated under yardstick competition. Rather than individually assessing prices and imposing productivity targets based on regulatory estimates, targets are now set collectively and on the basis of actually observed average costs and performance levels. Firms that are able to perform better than the average, generate additional profits while those who do not, earn less profits. Thus, artificial competition is created: Firms have a continuous incentive to operate more efficiently and provide an optimal quality level than others. At the same time, this leads to an even higher average performance of the industry as a whole.

The application of yardstick competition by DTE is, however, different from the classic textbook approach by Shleifer (1985). According to Shleifer´s original proposal, the yardstick for a given firm is set at the average costs of all other firms. In the Dutch case, prices are adjusted by an X-factor that reflects the change in the total factor productivity of efficient (frontier) firms. Thus, the yardstick formula is one whereby prices are based on relative changes in productivity rather than on absolute costs. The same principle is also applied to quality. A more detailed description of the price and quality parts of the Dutch approach follows in the next sections.

Price Regulation[1]

For each firm, an annual allowed revenue is determined. Based on this, tariffs are set in such a way that the allowed revenue equals the sum of revenues generated by these tariffs and corresponding volumes. These volumes, denoted as y_{2000}, act as so-called “norm volumes”. The calculation is made in the tariff basket where, for each year, prices are set in such a way that a firm i is allowed to earn the following revenue (DTE 2004a):

$$AR_t = \sum p_{j,t} \cdot y_{2000,j,t} \qquad \text{(VI-1)}$$

AR stands for Allowed Revenue and results from the yardstick scheme; p_j stands for the j^{th} tariff/price component. This formula is the starting point for setting annual tariffs. AR comes from an opening total cost benchmarking and is set on the basis of the yardstick scheme, following the general CPI-X formula:

$$AR_t = (1-X)^t \cdot AR_{2003} \tag{VI-2}$$

So far, the system can be described as a common tariff basket system. The main issue that sets this system apart from others is the way in which the X-factor is set. This is done according to the following formula:

$$(1-X_i)^3 = \theta_i \cdot (1-g)^3 \tag{VI-3}$$

Here, θ is the efficiency score for each particular firm and this value directly follows, from DTE´s 2000 totex DEA study bundling opex and capex efficiency analysis together. The factor *g* is the "frontier shift", measured on the basis of the change in total factor productivity for frontier firms (those with an initial or corrected efficiency score of θ=100 percent). Note that the X-factor is firm specific, but the *g* one is the same for all firms. Thus, each firm has its own X-factor that consists of two separate components:

1. A general target (*g*), which is the same for all firms and is based on the change in total factor productivity during 2004-2006; and
2. An individual target (θ_i) which is set individually by firm, and aims at removing initial efficiency differences across firms.

On top of the above formulas, a CPI adjustment is applied (CPI-X). Note that for the purpose of measuring productivity changes and making any appropriate ex post corrections, costs need to be deflated to a common base year.

The general idea of the Dutch yardstick system is that prices are adjusted on the basis of realised changes in general productivity (i.e. the frontier shift). This is done by measuring the change in the frontier (as reflected in factor *g*). This factor *g* is determined only by those firms that are 100 percent efficient on a total (not just operating) cost basis, i.e. those that either initially have a total cost DEA score of 100 percent or otherwise have caught up to the frontier in the meantime. The reason for splitting the X-factor into two components is to recognise the fact that those firms that are initially inefficient should not be included in the calculation of the frontier shift. Thus, firms that have less than 100 percent efficiency will be given two targets, namely (1) catching up to the frontier; and (2) shifting with the frontier. After the second regulatory period, the value of θ will be set to one by default. Therefore, DTE will then assume that all firms are equally efficient (level playing field) and that no further adjustment is needed to the yardstick scheme. At that stage, the X-factor will be set on the basis of *g* only and is derived from the relative change in the productivity of frontier firms. Here, productivity means TFP, i.e. it is defined in terms of an index number consisting of a ratio of weighted outputs to weighted inputs. As inputs, DTE uses the following cost items:

- Operating costs (opex);
- Standardised depreciation (*depr*); and
- Return on standardised book values (WACC times the Regulatory Asset Base or RAB).

Inputs are weighted equally, i.e. the single input in the comparative efficiency exercise is simply given by:

$$totex = opex + depr + WACC \cdot RAB \tag{VI-4}$$

As an output, DTE uses a so-called "proxy" output which is defined as the weighted sum of sold quantities (measured by the number of consumers, kWh, kW, etc.) for all consumer classes. The weighting is made on the basis of predefined factors that are set by DTE. The weighting factors correspond to an average tariff basket for the industry. These weighting factors are fixed for the entire regulatory period.

In the totex efficiency analysis, a firm becomes efficient if its costs per proxy output in 2002 are less than or equal to the projected/target efficient level (which is the efficient, or frontier, cost level obtained under totex DEA). The general view of the regulator is that a firm will be labelled a 'frontier' firm if it managed to achieve the efficiency improvement required by its DEA score in 2000 by the end of the regulatory period. Any super-efficiency achievements will be pocketed, in the form of unregulated extra profits, by the over-achievers. Once a firm reaches the frontier, it will be considered fully efficient in the future by default. That is, all firms with a DEA score of 100 percent in the 2000 benchmark by definition will stay on the frontier – irrespective of their performance afterwards. Note that this is a policy assumption which might turn out to be wrong because of efficiency leapfrogging i.e. some initially inefficient firms may by then well have surpassed initially efficient firms.

The value of g is not known at the start of the period, but can only be measured once the productivity improvements have been achieved ex post. Therefore, DTE will have to forecast g at the start of the regulatory period. At the end of the period, DTE will measure the "true" g and will correct for any differences – including interest – in the subsequent regulatory period. It is worth noting that, at the start of the third regulatory period, DTE will not yet have full information about costs in the last year of the second period. Therefore, the comparison between the forecast and the actual improvements will be excluded this last year.

Quality regulation

DTE considers reliability of supply as the main quality dimension (DTE 2004b). Reliability is a measure for the ability of the network to continuously meet the demand from consumers. It is characterised by the number and duration of interruptions experienced by consumers. As part of the quality regulation, use will be made of the indicator SAIDI. The quality regulation only applies to faults occurring in the medium and low voltage distribution networks. Interruptions that have their underlying cause in the high voltage networks or generation plants are excluded from the regulation.

For the purpose of regulation, uniform performance data is an important precondition; non-uniform data can unjustly affect the outcomes of the regulation system. For collecting reliability performance information, use will be made of reliability statistics collected under the so-called Nestor project. In the Netherlands, reliability statistics have been collected since 1975 as part of the Nestor project, which is a joint collaboration of Dutch distribution firms. However, since the Nestor project was not incepted as a regulatory tool, its original objectives are slightly different from the regulatory requirements. The Nestor effort was primarily component driven and aimed at obtaining a better understanding of faults in different components and types of electricity networks. At the same time, within the Nestor project, there is already substantial experience in collecting reliability data. Over the past 25 years, the quality of the data collection increased, with improvements still ongoing. In order to enable the Nestor results to be used for the purpose of quality regulation, definitions and measurement procedures will need to be harmonised throughout the industry.

The regulation system for reliability can best be characterised as a continuous incentive system with both penalties and rewards. The level of the reward or penalty depends on the deviation from a quality target. If, for a given year, the actual level of reliability of a certain firm is higher than the target, the firm receives a reward. On the other hand, if actual reliability is lower than the target, the firm pays a penalty. The rewards and penalties come in the form of a revenue increase respectively decrease. This can be represented in a simplified form as:

$$\pi_{i,} = \varphi \cdot (SAIDI_{\mathrm{target}} - SAIDI_{actual}) \tag{VI-5}$$

Here, π stands for the adjustment in the revenue (penalty or reward) for firm *i* in year *t*. Two main components of the incentive can be identified. Firstly, the quality target which is the term $SAIDI_{target}$. At first sight, the Dutch quality incentive scheme is very similar to the Norwegian one. There is, however, an important difference. Under the Norwegian scheme, the quality target is fixed and does not change its value during the course of the regulatory period. In contrast, in the Dutch case, the quality target is set at the end of the regulatory period based on observed levels of average performance of the industry. By doing so, each firm has an incentive to outperform the other. In turn, this leads to a further improvement in the average quality level and therefore the yardstick. At the same time, each firm also has an incentive to provide an optimal quality level. Similar as under the Norwegian CENS system, the firm now needs to make a trade-off between its own costs (maintenance, investments, etc.) and the financial effect of its quality choice. Spending more leads to higher quality and therefore potentially a positive financial incentive i.e. quality reward. However, higher costs also reduce profits and a quality increase will only be profitable if the associated costs are lower than the anticipated increase in the quality incentive.

The second component of the quality formula is the marginal quality incentive φ, at which the difference between targeted and actual performance is priced. The basic idea is that – since reliability and costs are inextricably related - firms will choose a reliability level at which its total profits are maximised. By internalising the quality effects into the firm's profit, the firm will make a trade-off between costs and quality performance that will be driven by the value of φ. If the value for φ is chosen correctly, then the firm's private reliability level choice will equal the socially optimal one. That is, if the value for φ is set at the level that reflects consumer willingness to pay for reliability, a situation can be achieved where marginal costs equates the marginal consumer valuation for quality i.e. a socially optimal outcome can be attained (see also chapter four).

The value for φ has been determined on the basis of a survey conducted by the University of Amsterdam on behalf of DTE. The main survey results are shown in Table VI-1. A distinction is made between two consumer groups namely households and businesses. Furthermore, as can be observed, interruption costs are measured as a function of two variables namely the annual frequency of interruptions (*F*) and the average duration of these interruptions (*D*). Given that the quality indicator in Equation (VI-5) is SAIDI, the value for φ needs to be converted in terms of costs per minute of interruption. In doing so, DTE made an estimation of the total interruption costs experienced by Dutch consumers in the year 2003 and divided this by the total number of minutes of interruption in that year.

Table VI-1. Results of the interruption cost study by DTE. Source: SEO (2004).

Households	
F>0.12, D>0.35	$2.30 \cdot \ln(0.08 \cdot (1+100 \cdot F)) \cdot \ln(2.89 \cdot D)$
F≤0.12, D>0.35	$-10.3 \cdot (1-F) + 4.74 \cdot \ln(2.89 \cdot D) \cdot F$
F>0.12, D≤0.35	0
F>0.12, D≤0.35	$-10.3 \cdot (1-F)$
Businesses	
F>0.08, D>0.24	$15.4 \cdot \ln(0.11 \cdot (1+100 \cdot F)) \cdot \ln(4.19 \cdot D)$
F≤0.08, D>0.24	$-73.8 \cdot (1-F) + 36.5 \cdot \ln(4.19 \cdot D) \cdot F$
F>0.08, D≤0.24	0
F>0.08, D≤0.24	$-73.8 \cdot (1-F)$

Notes

[1] This section is based on Ajodhia et al. (2005).

Annex VII. Dataset DEA sample

	TOTEX [EUR]	SOTEX [EUR]	CML [MIN/YEAR]	CONSUMERS	ENERGY [kWH]
UK-1	205,930	215,898	87,511,914	2,083,617	27,008,000
UK-2	135,941	151,094	133,032,000	1,446,000	15,444,000
UK-3	130,037	140,358	90,610,000	1,066,000	12,643,000
UK-4	240,344	262,924	198,225,580	2,437,000	28,949,000
UK-5	163,014	178,330	134,464,000	1,528,000	16,974,000
UK-6	207,943	224,579	146,048,000	2,282,000	25,444,000
UK-7	155,669	164,411	76,744,000	1,448,000	16,756,000
UK-8	181,302	205,014	208,162,000	2,146,000	21,154,000
UK-9	209,527	223,035	118,580,000	2,156,000	24,268,000
NL-1	12,657	12,960	2,511,052	118,390	1,163,000
NL-2	119,337	122,070	22,581,501	888,336	10,256,479
NL-3	72,705	74,703	16,514,110	493,843	4,647,330
NL-4	121,203	128,010	56,252,258	1,062,165	11,998,000
NL-5	8,889	9,725	6,909,786	99,450	886,933
NL-6	1,470	1,550	661,807	11,755	71,604
NL-7	4,682	4,719	311,738	45,114	330,741
NL-8	4,813	4,821	69,444	45,388	465,836
NL-9	80,146	80,782	5,262,292	456,400	9,074,018
NL-10	8,821	9,031	1,738,392	86,059	868,105
NL-11	129,982	130,926	7,797,846	938,369	13,569,000

Annex VIII. Binary Encoding of Networks

The number of MV substations can be one, two, or four. Based on the number of MV substations, the supply area is divided into equally sized sub-areas with the MV substation located in the centre of each sub-area. Each MV substation supplies loads located in the corresponding sub-area only, there are no electrical connections between different substations or sub-areas. The number of secondary feeders per MV substation can take any value up to 31. Once the number of secondary feeders has been determined, the set of MV/LV transformers to be fed from each MV substation is distributed among the different feeders. Each load point is assigned to a certain feeder.

If a MV substation is supplied via a single primary feeder, a fault in the primary feeder leads to an immediate cessation of the supply to all consumers connected to that MV substation. A second primary feeder (as well as related protection systems) adds redundancy to the system and increases quality. In that case, a fault in one of the primary feeder no longer leads to an interruption as then, the faulted primary feeder is automatically isolated, and supply is maintained via the remaining healthy primary feeder (assuming that it has sufficient capacity).

To identify the possible routing of the feeders, first a Delaunay triangulation of each set of MV/LV transformers and the corresponding MV substation is performed.[1] The Delaunay triangulation produces a set of edges that form feeder paths suitable for connecting the transformer stations to the corresponding substation (Peco and Gómez 2000). Figure VIII-1 shows the Delaunay triangulation of the supply area studied in chapter six.

Based on the identified feeder paths, an initial routing of the network can be performed using either the Dijkstra or the Travelling Salesman algorithm.[2] Starting from a given MV substation and for each feeder connected to the substation, the Dijkstra algorithm connects all the loads assigned to that feeder in such a way that the total length of the feeder is minimised. This process is repeated for each MV substation. On the other hand, the Travelling Salesman algorithm determines the shortest closed loop that joins a number of load points. The starting and end points of each loop belong to feeders of the same substation.

Secondary feeders can be radial or meshed. If the feeder is meshed, loads connected by that feeder can be supplied via two or more possible routes. In case of an interruption in the standard feeder route, supply can be restored via an alternative route. Firstly, the faulted part of the feeder is isolated and then switching actions are performed in such a way that an alternative feeder route is created from the MV substation to the interrupted load. Once isolated, the faulted components of the feeder can be repaired with supply to consumers already being restored. This reduces the average duration of interruptions compared to a radial feeder where supply can only be restored once the faulted component has been repaired. Here, repair may either comprise physical repair of the component or replacing it by a new healthy one.

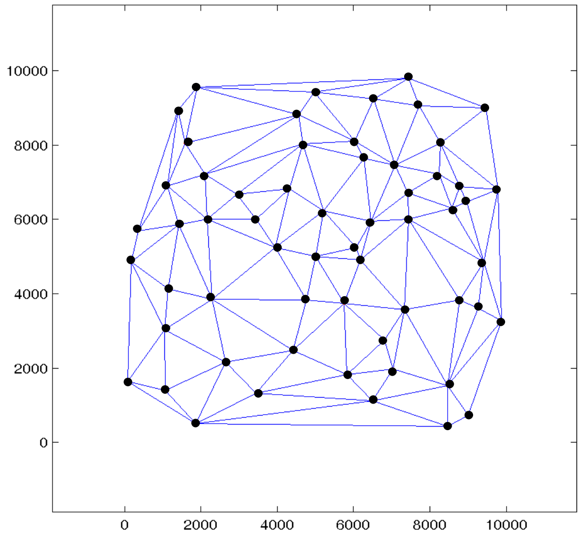


Figure VIII-1. Delaunay triangulation of the supply area from Figure 6-4.

The capacity of each feeder is selected from a list of available feeder types. This selection depends on two factors. Firstly, each feeder should have a minimal capacity in such a way that the system can serve the peak load under normal circumstances i.e. in a situation without any faults. In the case of a meshed network, when a failure occurs, it may be that certain feeders are loaded beyond their capacity as a result of rerouted power flows. If such circumstances occur, then the feeder capacity needs to be increased subsequently. Feeders may furthermore be protected or not. Additional feeder protection limits the impact of a fault to the respective protection zone.

Notes

[1] For a detailed discussion of Delaunay triangulation, see for example Lee and Schachter (1980).

[2] For a detailed description of the Dijkstra and Travelling Salesman algorithms, see for example Johnsonbaugh (1993).

Annex IX. Branch Replacement Algorithms

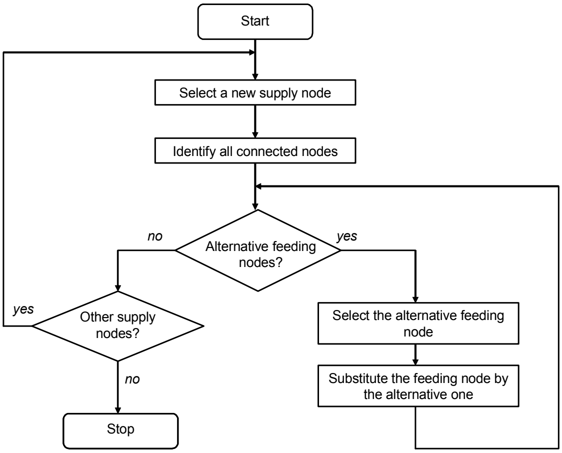


Figure IX-1. Algorithm for Branch Replacement Type I.

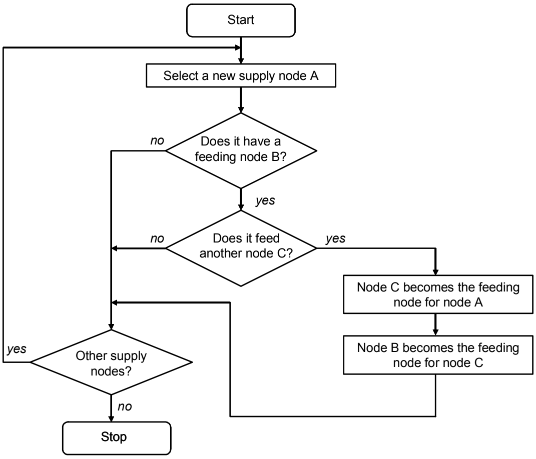


Figure IX-2. Algorithm for Branch Replacement Type II.

Summary

Regulating Beyond Price

Integrated Price-Quality Regulation for Electricity Distribution Networks

V.S. Ajodhia

ISBN-10: 9090196528
ISBN-13: 9789090196527

Presently, price-cap regulation is widely being applied in network industries such as electricity, gas, telecom and water. The main advantage of a price-cap system is its incentives for higher productive efficiency. There is, however, a growing concern that price-caps may also result in problems at the quality front as the firm may attain part of the cost savings through an undesired reduction in quality. This thesis aims to develop an integrated approach for optimal price-quality regulation of electricity distribution networks under a system of price-caps. In doing so, an integrated view of price and quality is taken.

Under rate-of-return regulation, the firm has a tendency to overcapitalise. This consequently leads to high quality levels. In contrast, price-cap regulation has superior efficiency properties but is also associated with an explicit need for integrating quality into the price-cap. This, however, is a complex undertaking suggesting that at some point, the advantages of stricter price regulation – coming from an enhancement of efficiency levels – will not outweigh the additional regulatory costs of setting in place adequate quality regulation.

Price-caps unlink prices from actual costs by imposing predefined change in prices over the course of a fixed regulatory period. The annual change in prices is determined by the X-factor. Two main price-cap approaches can be identified. Firstly, under the building blocks approach, operational and capital expenditures (opex and capex) are regulated on a separate basis. Secondly, under the total cost or totex approach, the regulator makes no distinction between opex and capex.

Under both price-cap approaches, benchmarking is an important regulatory instrument in estimating the X-factor i.e. the firm's potential for improvement. Integrating quality into the price-cap essentially takes the role of integrating quality into the benchmarking analysis. Under

building blocks, integration comes in the form of conducting an integrated price-quality assessment of the firm's proposed investments. Here, the regulator would need to simultaneously assess whether the implied investment leads to a desirable level of quality, and whether the investment is undertaken at least costs. For the totex approach, quality integration implies the application of an integrated price-quality benchmarking analysis of the firm's actually incurred total costs and supplied quality levels.

It seems that the quality problem is more severe under totex than under building blocks. Under the former approach, the regulator leaves the discretion of deciding on investments with the firm and subsequently cannot influence the quality level that the firm is providing. Under building blocks on the other hand, the regulator can indirectly steer quality levels by prescribing the desired investment level.

Determining the desired investment target is, however, difficult. This observation provides insight in the trade-off between price and quality. Under building blocks, quality degradation can be avoided by allowing a high investment level. This leads to higher prices and is essentially the premium for assuring high quality. Thus, the trade-off is biased towards quality rather than price. Under totex on the other hand, the regulator cannot directly control the firm's investment level and therefore not influence quality. On the other hand, efficiency incentives are higher as the firm has an incentive to reduce costs. Thus, under totex, the price-quality trade-off is somewhat biased towards price.

One approach to regulate quality is by choosing a partially integrated approach. Two classes of quality controls can then be distinguished. Firstly, indirect quality controls aim to provide consumers with information about the firm's quality performance and create institutions through which these better-informed consumers can demand or pressurise the firm to deliver an appropriate quality level. Under indirect controls, the role of the regulator is primarily one of an information provider and facilitator of disagreements on quality between firms and consumers. The second class of quality controls are direct controls. Here, the regulator provides the firm with direct financial incentives (penalties or rewards) in order to provide an appropriate quality level. Such direct controls come in the form of minimum standards or incentive schemes. Here, the regulator plays an active role; he develops a view of what quality levels to aim at and provides the firm with incentives to reach these.

Next to a partially integrated approach, the regulator could opt for a fully integrated price-quality approach. Here, the regulator sets the X-factor directly as a function of quality. If the regulator is better informed about the firm's price and quality performance, he can establish more effective incentives i.e. assure that the firm operates in a socially optimal way as well as assure that consumers share in the benefits resulting from the increase in social welfare. Benchmarking is an important regulatory asset in acquiring information about the firm's performance. Traditionally, benchmarking under price-caps has been mostly restricted to the analysis in the area of costs. A truly integrated benchmarking approach of price and quality was, however, lacking. This thesis has developed integrated price-quality benchmarking methodologies for the respective regulatory approaches.

For totex, a methodology for integrating quality into a Date Envelopment Analysis (DEA) framework is developed. Two basic approaches exist to incorporate quality into a DEA model. At the one hand, under a technical or model, quality can be defined as a technical output factor. On the other hand, the sotex model uses total social costs as an input factor. The application of the two DEA models to the sample of Dutch and UK firms provides an evaluation of the two methods. The empirical analysis shows that under the technical model, firms that specialise in cost or quality may be incorrectly classified as efficient. At the same time, the limitations of the sotex model were also illustrated. The sotex efficiency score may well potentially underestimate the true improvement potential in both the productivity as cost sense.

For building blocks, a benchmarking methodology was developed and implemented in the form of a software tool named the Network Simulation Tool (NST). The NST is based on the idea of comparing the performance of the firm's investment proposal to that of a large number of artificially constructed alternatives. These alternatives are generated through simulation and represent possible solutions that the firm might have considered instead of the one being proposed. The performance of an investment is measured in terms of its total social costs (sotex), which is defined as the sum of network costs and interruption cost.

The results of a case study using a fictive newly to be supplied are to indicate that the NST may well be an important tool in the process of investment appraisal. The NST can provide the regulator with valuable information regarding the preferred network characteristics for the given supply area. What is more, it allows the regulator to analyse the impact of changes in input parameters and supply area characteristics on the price-quality trade-off. Taking into account modelling and data restrictions however, the NST should be considered a tool aimed at revealing information about a range of investments and the desirable properties of a network rather than providing an exact prescription of what should be the optimal network for a given supply area.

An important conclusion of this thesis is that although benchmarking can play an important role in the determination of the price-quality integrated X-factor, the limitations of benchmarking should also be realised. If the regulator cannot properly measure costs, quality, or consumer demand for quality, the results of the benchmarking analysis can be adversely affected. On the other hand, the benchmarking outcome is driven by modelling assumptions that are bound to be imperfect. For benchmarking to remain effective, these limitations should be recognised and properly dealt with in the determination of the integrated X-factor. The role of benchmarking tools therefore is primarily one of assisting the regulator in extracting (albeit imperfect) information from the firm in order to allow him to estimate the range within which the X-factor should fall. The better the regulator is able to conduct the benchmarking analysis, the less broad this range will be and the more accurately the X-factor can be set. Benchmarking thus is not the ultimate solution, but rather a tool that one could use to better deal with the regulatory information problem.

Samenvatting

Regulating Beyond Price
Integrated Price-Quality Regulation for Electricity Distribution Networks
V.S. Ajodhia

ISBN-10: 9090196528
ISBN-13: 9789090196527

Price-cap systemen worden op wijde schaal toegepast voor het reguleren van netwerk industrieën zoals elektriciteit, gas, telecommunicatie en water. Het belangrijkste voordeel van een price-cap systeem is dat het leidt tot hogere productiviteit. Er is echter een toenemende bezorgdheid dat price-caps ook kunnen leiden tot problemen op het gebied van de kwaliteit. Het gereguleerde bedrijf kan mogelijk een kostenbesparing doorvoeren ten koste van een ongewenste afname in de kwaliteit. Het doel van dit proefschrift is om een aanpak te ontwikkelen voor het geïntegreerd reguleren van de prijs en kwaliteit van elektriciteitdistributienetwerken onder een systeem van price-caps. Hierbij worden prijs en kwaliteit vanuit een geïntegreerde optiek benaderd.

Onder rate-of-return regulering neigt het bedrijf naar overinvesteringen. Dit heeft een hoog kwaliteitsniveau tot gevolg. Price-cap regulering aan de andere kant heeft superieure eigenschappen voor wat betreft efficiëntie maar vergt tegelijkertijd een expliciete regulering van kwaliteit. Dit is echter complex om te bewerkstelligen hetgeen suggereert dat de voordelen van striktere vormen van prijsregulering – als gevolg van het efficiëntievoordeel – op een bepaald moment niet meer opwegen tegen de kosten van het additioneel reguleren van kwaliteit.

Price-cap regulering verbreekt de relatie tussen werkelijke kosten en de prijs door het vooraf bepalen van de prijs voor een gegeven reguleringsperiode. De jaarlijkse aanpassing in de prijs wordt weergegeven door de X-factor. Er bestaan twee soorten van price-cap systemen. Ten eerste, de building blocks benadering waarbij operationele kosten en kapitaalkosten onafhankelijk van elkaar worden gereguleerd. Ten tweede, de totale kosten of totex benadering waarbij er geen onderscheid wordt gemaakt tussen deze twee kostensoorten.

Het gebruik van benchmarking is een belangrijk regulatorische instrument onder beide (deze twee) price-cap benaderingen. Benchmarken wordt gebruikt om de X-factor te bepalen; deze representeert de door de regulator geanticipeerde verbetering in de efficiëntie van het bedrijf. Het integreren van kwaliteit in de price-cap komt dan feitelijk neer op het integreren van kwaliteit in de benchmark analyse.

In het geval van building blocks betekent integratie het uitvoeren van een geïntegreerde prijs- en kwaliteitsanalyse van de door het bedrijf voorgestelde investeringen. Hier zal de regulator dienen na te gaan of de voorgestelde investeringen leiden tot een gewenst kwaliteitsniveau en of deze investeringen tegen zo laag mogelijke kosten worden uitgevoerd. Voor wat betreft de totex benadering, betekent integratie van kwaliteit het uitvoeren van een gecombineerde benchmark analyse van de werkelijke door het bedrijf ondervonden totale kosten en geleverd kwaliteitsniveau.

Het schijnt dat het kwaliteitsprobleem erger is in het geval van de totex aanpak dan onder de building blocks aanpak. In het eerste geval heeft het bedrijf in principe de vrijheid zelf het gewenste investeringsniveau te kiezen. De regulator kan geen invloed uitoefenen op het door het bedrijf aangeboden kwaliteitsniveau. In tegenstelling hiermee kan de regulator kwaliteit indirekt bepalen in het geval van de building blocks aanpak. In dit geval bepaalt de regulator het gewenste investeringsniveau en daardoor ook indirect het kwaliteitsniveau.

Het bepalen van het gewenste investeringsniveau is echter een moeilijke opgave. Deze observatie biedt inzicht in de afweging die er bestaat tussen prijs en kwaliteit. In het geval van building blocks kan kwaliteitsdegradatie worden voorkomen door een hoger investeringsniveau dan strikt noodzakelijk toe te staan. Dit leidt tot hogere prijzen maar is tevens de premie voor een hoog kwaliteitsniveau. In dit geval is de afweging in het voordeel van kwaliteit. In het geval van totex is de situatie omgekeerd. Hier kan de regulator het investeringsniveau en daardoor het te leveren kwaliteitsniveau niet beïnvloeden. Er zijn echter meer voordelen op het gebied van efficiëntie aangezien het bedrijf sterkere prikkels heeft om zijn kosten te reduceren

Een manier om kwaliteit in de prijsregulering te integreren is de zgn. gedeeltelijke aanpak. In dit geval kunnen twee soorten van systemen voor het reguleren van kwaliteit worden geïdentificeerd. Ten eerste, indirecte systemen welke ten doel hebben consumenten van informatie te voorzien over de kwaliteitsprestaties van het bedrijf en instituties te creëren door middel welke consumenten eisen kunnen stellen of druk kunnen uitoefenen op bedrijven om een geschikt kwaliteitsniveau aan te bieden. De tweede mogelijkheid om kwaliteit te reguleren is via de zgn. directe systemen. In dit geval past de regulator financiële prikkels toe (boetes en beloningen) om zodoende het bedrijf te motiveren een geschikt kwaliteitsniveau aan te bieden. Deze systemen komen in de vorm van minimum standaarden en incentive systemen.

In het geval van indirecte systemen is de rol van de regulator passief. Deze is beperkt tot informatievoorziening en het bieden van faciliteiten om meningverschillen tussen consumenten en bedrijven op te lossen. In het geval van directe systemen speelt de regulator een actieve rol; hij ontwikkelt een idee van een geschikt kwaliteitsniveau en biedt het bedrijf prikkels aan om deze te bereiken.

Naast een gedeeltelijke aanpak is het ook mogelijk te kiezen voor een volledig geïntegreerde benadering van prijs en kwaliteit. In dit geval wordt de X-factor bepaald mede als functie van kwaliteit. Indien de regulator over meer informatie beschikt, is hij beter in staat om het bedrijf van effectievere prikkels te voorzien om een optimaal kwaliteitsniveau aan te bieden evenals om consumenten te laten meedelen in de toename in de sociale welvaart. Benchmarking is een belangrijk instrument om informatie te vergaren over de prestaties van het bedrijf. Bestaande benchmark analyses zijn echter beperkt tot de kosten. Een volledig geïntegreerde analyse van prijs en kwaliteit was tot nu niet beschikbaar. In dit proefschrift worden twee volledig geïntegreerde benchmarking methodes ontwikkeld voor elk van de price-cap benaderingen.

In het geval van de totex benadering is een methode ontwikkeld om kwaliteit te integreren in het Data Envelopment Analysis (DEA) model. Er bestaan twee manieren om dit te bewerkstelligen. De eerste manier, genoemd het technisch model, is om kwaliteit te definiëren als een output factor. Bij de tweede manier, het sotex model, wordt kwaliteit gedefinieerd als een input factor en wordt er gebruik gemaakt van het concept van totale maatschappelijke kosten. De twee modellen zijn geëvalueerd aan de hand van een toepassing van deze modellen op een dataset van Nederlandse en Britse bedrijven. De empirische analyse wijst uit dat in het geval van het technisch model, bedrijven die zich specialiseren in kosten of kwaliteit onterecht gekwalificeerd kunnen worden als te zijn efficiënt. Ook de tekortkomingen van het sotex model zijn gedemonstreerd. Het blijk dat de sotex efficiëntie score het werkelijke potentieel voor verbetering op het gebied van zowel de prijs als kwaliteit onderschat.

Voor het geval van de building blocks benadering is ook een benchmark methode ontwikkeld. Deze is geïmplementeerd in de vorm van een software programma genaamd de Network Simulation Tool (NST). De NST is gebaseerd op de idee de prestatie van een door het bedrijf voorgestelde investering te vergelijken met een groot aantal andere artificieel geconstrueerde alternatieven. Deze alternatieven worden gegenereerd door middel van simulatie en representeren andere mogelijke oplossingen naast die door het bedrijf voorgestelde investering. De prestatie van een alternatief wordt gemeten aan de hand van de daaraan gerelateerde maatschappelijke kosten (sotex). Sociale kosten zijn gedefinieerd als de som van de netwerkkosten en de kosten die consumenten ondervinden als gevolg van onderbrekingen.

De resultaten van een case studie tonen aan dat de NST wellicht een belangrijke rol kan spelen bij het proces van evalueren en het toestaan van investeringen. De NST kan belangrijke informatie beschikbaar stellen m.b.t. de vereiste karakteristieken van een investering. Tevens biedt de NST de mogelijkheid de invloed van input parameters en eigenschappen van het te voorzien gebied op de wenselijkheden van de investering te analyseren. De beperkingen in de modellering en de data dienen daarbij echter in acht te worden genomen. De NST geeft slechts informatie over de globale karakteristieken van een effectieve investering; het schrijft echter niet voor hoe deze investering er in detail uit zou moeten zien.

Een belangrijke conclusie van dit proefschrift is dat benchmarking een belangrijke rol kan spelen bij het bepalen van een geïntegreerde X-factor voor prijs en kwaliteit. Echter, de beperkingen van benchmarking dienen ook te worden erkend. Indien de regulator niet in staat is om kosten, kwaliteit, en consumenten vraag naar kwaliteit te meten, dan kunnen de resultaten van de benchmark analyse ongewenst worden beïnvloed. Aan de andere kant is de uitkomst van de benchmark analyse mede bepaald door de aannames in de modellering – deze aannames zijn in principe imperfect.

Om benchmarking effectief te laten zijn, is het belangrijk de genoemde beperkingen onder ogen te zien en mee te nemen bij het bepalen van de X-factor. De rol van benchmarking is derhalve voornamelijk die van het extraheren van (imperfecte) informatie om zodoende een geschikte afbakening van de X-factor te bewerkstelligen. Des te beter de regulator in staat is de benchmark analyse uit te voeren, des te beter kan de X-factor worden afgebakend. Benchmarken is daarom niet de ultieme oplossing maar meer een belangrijk instrument welke gebruikt kan worden voor het omgaan met het probleem van de regulatorische informatie asymmetrie.

Curriculum Vitae

Virendra Shailesh Ajodhia was born on the 18th of June 1975. He completed the Mr. Dr. J.C. de Miranda Lyceum in Paramaribo, Suriname (1993) and studied Electrical Engineering at Delft University of Technology, Delft. After graduation in 1999, he joined KEMA Consulting in Maastricht, the Netherlands. From 2001 to 2002, he was a regulatory economist for the Dutch Energy Regulator (DTE) in The Hague, the Netherlands. From September 2002 he operates as an independent consultant and is associated with KEMA Consulting in Bonn, Germany. He has provided consulting services to numerous utilities and regulators in Europe, Asia, and the Caribbean. Currently he is based in Paramaribo, Suriname. In 1999, he married Shashi Soerjoesing; they have one daughter, Juhi (2005).