Cost Efficiency and Regulation of Slovenian Water Distribution Utilities: an Application of Stochastic Frontier Methods

Presented by
Jelena Zoric

Thesis supervisor: Prof. Massimo Filippini
President: Prof. Rico Maggi
External examiner: Prof. Nevenka Hrovatin

A dissertation submitted to the
Faculty of Economics
University of Lugano
Switzerland

For the degree of
Ph.D. in Economics

Lugano, October 2006

Acknowledgements

I am deeply grateful to my supervisor, Professor Massimo Filippini, for his guidance, comments, suggestions and fruitful collaboration. His advice throughout all stages of the thesis was invaluable. By observing his approach to research I have learned a lot. I am especially thankful for his suggestions regarding the empirical part of this work. Special thanks go to the MecoP Institute at the University of Lugano and all of its members for providing the facilities and a stimulating environment for the research.

Very special thanks are due to Professor Nevenka Hrovatin from the Faculty of Economics in Ljubljana for introducing me to the field of the regulation and efficiency of public sector utilities. Without her support I would probably have never started my research work in this field.

I would like to thank the Ministry of the Environment and Spatial Planning of the Republic of Slovenia, especially Majda Podlipnik for her help in issuing the questionnaire and collecting the data for the empirical study.

I am very grateful to the Faculty of Economics in Ljubljana for providing substantial funding during my PhD studies at the University of Lugano. I also owe gratitude to the Fondazione Daccò, the Bank of Slovenia and the Ad futura, Science and Education Foundation of the Republic of Slovenia for providing additional funding for my studies.

Finally, I wish to thank my family and friends for their encouragement and support throughout my doctoral studies.

Abstract

Over the last two decades it has become increasingly important to promote the efficiency and improve the performance of natural monopolies operating within network industries. In line with this, different regulatory approaches have been designed aiming at preventing the abuse of monopoly power and at the same time enhancing the performance of regulated firms. The most widely adopted incentive-based regulatory schemes involve price cap (RPI-X), revenue cap, and yardstick regulation models. These schemes aim to give firms an incentive for efficient production and cost reduction. However, due to the imperfect information available to the regulator there are some drawbacks related to the use of price-cap regulation since the regulator does not know a firm's true costs. High costs may be due to a firm's particular production situation or merely to its inefficiency. Thus, in setting the initial price level and the yearly efficiency factor X in the price-cap formula the regulator can use some form of cost-based benchmarking analysis. In this case, benchmarking is used to establish a larger information basis for more effective regulation that reduces the informational asymmetry between firms and the regulator. Hence, there is a close link between efficiency measurements and incentive-based price regulation.

Today's price regulation of Slovenian water distribution utilities resembles the rate-of-return regulation scheme. Nevertheless, the current Rules on Price Determination of Obligatory Local Public Utilities for Environmental Protection (2004) envisage the use of benchmarking in the price-regulation process and defining of the best-practice performance. However, the Rules have not yet been put into practice since the benchmarking method has still not been determined. In the thesis we consider the use of parametric frontier benchmarking methods and suggest how the results could be used in the price-regulation process. The main method employed is Stochastic Frontier Analysis (SFA), while the Corrected Ordinary Least Squares (COLS) method is used to cross-check the results. A translog frontier cost function is estimated based on an unbalanced panel data set of 52 utilities over the 1997-2003 period. The cost inefficiency estimates of Slovenian water distribution utilities are obtained by several different parametric frontier methods. The employed models differ in their assumptions, method of estimation and in their ability to account for firm-specific effects and distinguish between firm heterogeneity and inefficiency. The pooled model does not take into account the panel structure of the data and is therefore unable to separate unobserved heterogeneity from inefficiency. While conventional fixed and random effects panel data models take firm-specific effects into account in the estimation of inefficiency, they treat any time-invariant unobserved heterogeneity as inefficiency. They, too, are found to fail when it comes to separating heterogeneity from inefficiency. This problem is tackled by 'true' fixed

and random effects models by adding in an additional term in the model which captures time-invariant and firm-specific effects and therefore separates these effects from inefficiency (Greene, 2002a, b). However, it remains debatable whether the time-invariant firm-specific effect should, in fact, be attributed to unobserved heterogeneity or to inefficiency. Mundlak's (1978) formulation of the random effects model is also considered since it allows controlling for any correlation between unobserved heterogeneity and regressors. In our study it is found that the estimation results based on the conventional random effects models tend to highly overestimate cost inefficiency, while the true fixed effects (TFE) model seems to slightly underestimate it. Nevertheless, since the inefficiency estimates obtained by the TFE model closely correspond to the pooled model it is believed that these two models provide a better approximation of the actual cost inefficiency of Slovenian water distribution utilities, which is found to be close to or slightly above 20% on average. The TFE model is also found to perform the best with respect to the expected signs and significance of the regression coefficients. The inefficiency results indicate that significant cost inefficiency is present in Slovenian water distribution companies and that the utilities would have to considerably cut their costs in order to become efficient. This may be facilitated by a properly designed price regulation that introduces incentives for efficiency improvements. The inefficiency scores obtained from the different methods are, however, not found to be consistent in their levels and rankings of the utilities. A possible explanation of these inconsistent results can be found in the different ability of stochastic frontier methods to account for unobservable heterogeneity. However, since the regulator needs reliable estimates of the efficiency potential of a regulated firm this finding is particularly unwelcome. It is thus recommended to use the benchmarking results obtained by the SFA methods only as a starting point for providing information about the range in which the inefficiency score can be located. Alternatively, the estimated cost function can also be used to predict utilities' costs, with this approach being in line with yardstick competition.

Besides achieving cost efficiency, i.e. operating at minimum cost at a given size, important cost savings may result from achieving scale efficiency, i.e. operating at the size that minimises average production costs. The results from the different models in the latter case prove to be fairly consistent. Based on the obtained results, the presence of economies of output density and customer density in Slovenian water distribution utilities is confirmed. Therefore, it would be beneficial for the utilities if they managed to distribute larger volumes of output to their existing customers as well as to acquire new customers. With respect to economies of scale, medium-sized utilities are found to closely correspond to the optimal size of water distribution utilities in Slovenia. Economies of scale prevail in smaller utilities, implying they should consider expanding the scale of their operations through mergers.

Conversely, large utilities are found to operate at levels where economies of scale are already exhausted.

Overall, based on the results obtained it can be concluded that there is large potential for cost savings in Slovenian water distribution utilities. However, no evidence of any notable improvements being made can be found so far. The total factor productivity growth over the examined period is found to be around zero where technical progress is established but, on the other hand, no significant improvements in cost efficiency are found. In order to facilitate these improvements, a new regulatory framework is needed, where the choice should be made among incentive-based price regulation schemes. Rate of return regulation combined with benchmarking as proposed by the Rules on Price Determination of Obligatory Local Public Utilities for Environment Protection (2004) would be one of the appropriate alternatives.

Keywords: *SFA, cost frontier function, cost inefficiency, price regulation, water distribution utilities*

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Background and the Study Problem

Network industries (water industry, electricity, gas, telecommunications, railways etc.) used to be typically vertically integrated national or regional monopolies with exclusive rights to serve customers. One reason for this industry structure was the common assumption that one single firm is able to operate with lower costs than if several firms were to supply the same level of output. Nonetheless, since the late 1980s a wave of reform has transformed the institutional framework, organisation and operating environment of network industries. Privatisation along with the market liberalisation of utilities and network infrastructure have become important policy objectives in many developed and developing countries. Although the structure of sectors and the approaches to reform vary between countries and sectors, the main aim is to improve the sector's efficiency, effectiveness and competitiveness which should, in turn, result in lower prices for final customers. For instance, due to technological progress and public pressure to decrease prices, most electric power reforms have focused on the introduction of competition in generation and supply, whereas due to their natural monopoly character transmission and distribution activities remain regulated (Jamasb and Pollitt, 2001). On the other hand, reforms of the water industry usually only involve the introduction of new incentive-based regulatory approaches since the technology does not facilitate the introduction of direct or side-by-side competition. Moreover, in some countries privatisation of the water industry has been carried out, for example, in the UK.

Despite the ongoing reforms, there is clearly still a need to regulate the water sector since it is characterised as a natural monopoly. Moreover, the regulation should be designed in such a way as to provide incentives for cost reduction and more efficient production. Regulatory authorities world-wide have adopted a variety of approaches to regulate distribution prices. The most widely adopted incentive-based schemes involve price cap (RPI-X), revenue cap, yardstick regulation and other benchmarking methods.[1] These schemes aim to give firms an incentive for efficient production and cost reduction. However, due to the imperfect information available to the regulator there are some problems with price-cap regulation since the regulator does not know a firm's true costs. High costs may reflect a firm's particular production situation or simply its inefficiency. Thus, if price caps are set too high

[1] A review of different regulatory schemes can be found in Joskow and Schmalensee (1986).

there is a possibility of a welfare loss while very low price caps could see firms ending up with viability problems. To overcome this informational asymmetry problem between the regulator and firms, some form of benchmarking analysis can be applied in setting productivity or efficiency requirements for regulated firms. In this case, benchmarking analysis is used to establish a larger information basis for more effective price-cap regulation. There is thus a close link between the efficiency measurement and incentive-based price regulation methods such as price cap and yardstick regulation. As will be discussed later, in some countries the regulatory authorities make direct use of benchmarking results in the process of setting prices.

Unfortunately, the evidence from empirical studies shows that the various benchmarking methods often produce different results with respect to firms' efficiency scores and rankings.[2] A possible explanation of this inconsistency problem relates to the difficulty of benchmarking methods in accounting for unobservable heterogeneity in environmental and network characteristics across companies (Filippini, Farsi and Fetz, 2005). This is particularly undesirable if the results are to be used in economic policy-making. The regulatory authorities are typically faced with a problem of choosing the most appropriate technique to be put into practice. Despite extensive research carried out in the field of efficiency measurement, so far there is no consensus on which method has been found to perform the best. Since the various benchmarking methods may lead us to different results and none of the methods has been proven to be superior with respect to the others, it is important to be aware of the advantages and disadvantages of applying the different benchmarking approaches to measure a firm's performance. In addition, it is important to study the consistency and reliability of the results obtained regarding firms' productivity or efficiency. In the absence of any consensus on the most appropriate technique to use, a purely pragmatic approach would entail a combination of results from the different models.

The current Slovenian price regulation of water utilities closely resembles the traditional rate-of-return regulation, which has not been shown to provide sufficient incentives for efficiency improvements. Nevertheless, it should be noted that the rules on price regulation recently issued by the government (i.e., Rules on Price Determination of Obligatory Local Public Utilities for Environment Protection, 2004) envisage the benchmarking of costs and quality combined with the rate-of-return regulation. However, the rules have not yet been put into practice, nor has been the benchmarking method specified. Incentive regulation like a price-cap scheme, yardstick competition or rate-of-return regulation combined with the

[2] For example, see Bauer et al. (1998), Estache, Rossi and Ruzzier (2004), and Farsi and Filippini (2004).

benchmarking of the costs appears to be a good alternative to be implemented in the Slovenian water sector. For example, since 2003 a price-cap regulation combined with benchmarking has been applied to Slovenian electricity distribution utilities (AERS, 2004). In the EU water industry context, the two best-practice examples are the UK regulator OFWAT and the Italian Regulation Authority where benchmarking combined with either a price-cap or rate-of-return regulation is already in use (OFWAT, 1999; Massarutto, 1999).

In the following thesis, the use of parametric frontier benchmarking methods is considered to analyse the performance of Slovenian water distribution utilities. Several different stochastic frontier methods are employed to estimate the cost frontier function and cost inefficiency of water distribution utilities. Since these utilities operate in different regions with different environmental and network characteristics that are only partially observed, it is essential to be able to distinguish between inefficiency and unobserved heterogeneity that influences the costs. Until recently, this issue has been neglected in empirical work since the traditional stochastic frontier models are unable to make a distinction between these two effects. As a result, heterogeneity has often been confounded with inefficiency. Since this may have serious financial consequences for regulated companies, it is crucial to be able to explicitly model cost differences that are due to heterogeneity and inefficiency. New developments in the field of stochastic frontier analysis, namely true random and true fixed effects proposed by Greene (2002a, b), can help us address this issue. Further, unobserved heterogeneity, if not properly accounted, might not only influence inefficiency estimates but might, if correlated with regressors, result in biased coefficients of the cost frontier function as well. To overcome this problem, Mundlak's (1978) formulation is considered. If the results based on the stochastic frontier analysis are supposed to be used by regulatory authorities then their reliability is vital. Thus, we analyse the consistency of the cost inefficiency estimates obtained from both conventional panel data models and the newly proposed models. We are especially interested in finding out whether accounting for unobserved heterogeneity in the model significantly influences the results. Finally, we propose how the results of benchmarking analysis could be employed in regulating water prices in Slovenia.

1.2 Goals

The main objective of the thesis is to estimate a cost frontier function for a panel data set of Slovenian water distribution utilities over the 1997-2003 period by using several parametric approaches in order to:

- estimate the cost inefficiency levels of water distribution companies considering the presence of unobserved heterogeneity in the model;

- analyse the reliability and consistency of the individual inefficiency estimates obtained by applying different parametric frontier benchmarking methods;
- establish the existence of economies of scale and density, and ascertain the optimal size of the water utilities;
- estimate and decompose the total factor productivity growth in the Slovenian water industry; and
- evaluate the relevance of the results obtained for economic policy-making and propose how the results could be used in price-regulation process.

1.3 Methodology

To place stochastic frontier analysis within the larger context of metric benchmarking methods and to justify the choice of the method, we briefly present different benchmarking methods and point out some of their main advantages and disadvantages. The benchmarking of utilities can broadly be defined as the comparison of some measure of actual performance against a reference or benchmark performance. It can be used in the incentive regulation to promote improved efficiency by rewarding good performance relative to some pre-defined benchmark. Since the rewards are based on performance, two key issues that emerge are the choice of appropriate benchmarks and the techniques used to measure the performance. According to the classification suggested by Jamasb and Pollitt (2001), actual performance can be measured against benchmarks that are derived from the 'best/frontier' practice or some 'representative/average' measure of performance. The classification of benchmarking methods is presented in Figure 1.1.

Average benchmarking methods may be used to mimic competition among firms with relatively similar costs or where there is a lack of sufficient data or firms with which to compare for the application of frontier methods. The regression-based average benchmarking method is the Ordinary Least Squares (OLS) method. OLS estimates the average production function or the cost function of a sample of firms which then serves as a benchmark in evaluating firms' performance. In the case of a panel data set, conventional fixed and random effects panel data models can also be applied. Another method based on average performance is the use of indices as the benchmark, such as Total Factor Productivity (TFP) index. Input quantities, output quantities and prices are required to construct a TPF index. An important advantage of the index-number approach is that it requires a minimal amount of data. It requires only two data points, either observations of two firms in one time period or observations of one firm in two periods, while the parametric approaches need a number of

firms to be observed. However, this approach does not account for noise and measurement error.

From a regulatory policy point of view, a big difference between average and frontier benchmarking is that the latter has a stronger focus on performance variations between firms. The frontier-based benchmarking methods identify or estimate the efficient performance frontier from the best practice used in an industry or a sample of firms. The efficient frontier then becomes a benchmark against which the relative performance of firms is measured, with inefficiency being viewed as a deviation from the optimal point on the frontier. Frontier methods can therefore be used for setting firm-specific efficiency requirements. This approach can be suitable in the initial stages of regulatory reform when a priority objective is to reduce the performance gap among utilities.

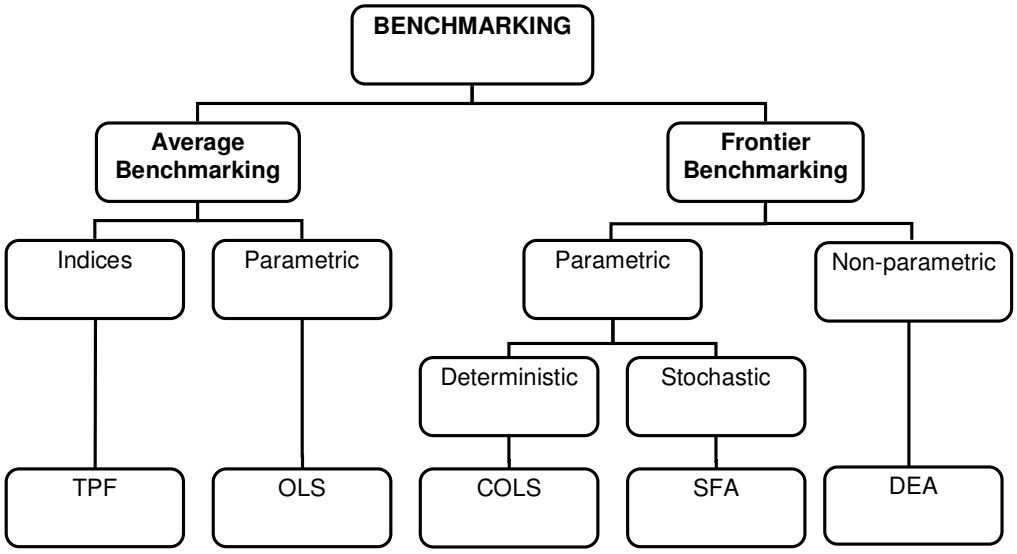

Figure 1.1: Classification of benchmarking methods according to Jamasb and Pollitt (2001)

The main frontier benchmarking methods are Data Envelopment Analysis (DEA), Corrected Ordinary Least Squares (COLS), and Stochastic Frontier Analysis (SFA). DEA is a non-parametric (linear programming) approach, while COLS and SFA are parametric (statistical) techniques.

DEA was introduced by Charnes, Cooper and Rhodes (1978) and Banker, Charnes and Cooper (1984).[3] The former paper assumes constant returns to scale, while in the latter variable returns to scale are assumed. DEA involves the use of linear programming methods to calculate (rather than estimate) a non-parametric piece-wise efficient frontier.[4] Firms that make up the frontier envelop the less efficient firms. Efficiency measures are calculated relative to this frontier. The efficiency of the firms is calculated in terms of scores on a scale from 0 to 1 with the frontier firms receiving a score of 1. DEA does not require the specification of a cost or production function, which is an advantage over the parametric methods. However, the method does not allow for stochastic factors and measurement errors, which can influence the shape and position of frontier. Outliers may also notably influence the results. Moreover, efficiency scores tend to be sensitive to the choice of input and output variables. The exclusion of important input or output data can lead to biased results. In addition, as more variables are included in the model the number of firms on the frontier increases.

Parametric frontier methods require the specification of a cost or production function and therefore involve assumptions about the technology of the firm's production process. Flexible functional forms are recommended in order not to impose overly restrictive assumptions on the technology. The COLS method is based on regression analysis and is relatively simple to implement. It was introduced by Winsten (1957). One drawback of the COLS method is that it does not allow for stochastic errors and relies heavily on the position of the single most efficient unit. It assumes that all deviations from the frontier can be attributed solely due to inefficiency. This shortcoming can be avoided by using stochastic frontier methods. In contrast with the deterministic COLS method, SFA recognises the possibility of stochastic errors. However, accounting for stochastic errors requires the specification of a probability function for the distribution of the statistical noise and inefficiencies. Compared to DEA, the main advantages of SFA are that it allows for stochastic error and that conventional tests of hypotheses regarding the existence of inefficiency and regarding the structure of the production technology can be conducted (Coelli, Rao and Battese, 1998).

SFA was introduced by Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977). In subsequent papers, Pitt and Lee (1981) and Schmidt and Sickles (1984)

[3] Good references for the theoretical aspects of DEA are Cooper, Seinford and Tone (2003) and Zhu (2003).

[4] Another non-parametric frontier technique to be mentioned is Free Disposal Hull (FDH). For details, see Cooper, Seinford and Tone (2003).

proposed stochastic frontier models for panel data. Over the years, many extensions to the originally proposed stochastic frontier models have been developed.[5] We will be in particular interested in those panel data models that are able to separate unobserved heterogeneity from inefficiency. Therefore, we investigate in some more detail the true fixed and true random effects models recently developed by Greene (2002a, 2002b), which capture the effects of time-invariant unobserved heterogeneity by a separate term. Mundlak's (1978) formulation of the random effects model as proposed by Farsi, Filippini and Kuenzle (2005) is also considered since this model has the ability to control for the correlation between unobserved heterogeneity and explanatory variables. Nevertheless, despite the intense research effort so far there is no consensus on what is the best frontier benchmarking method.

In the thesis we will focus on parametric frontier methods for estimating inefficiency and leave non-parametric methods aside. Arguments for this choice are the abovementioned shortcomings of DEA and the fact that parametric methods allow us to analyse in more detail not only the cost inefficiency but also the economies of size, output density and customer density. It should be also stressed that DEA is unable to account for unobserved heterogeneity, which may be quite high in network industries. As a result, the efficiency scores obtained may be considerably biased. Since one of the main focuses of the thesis is the separation of unobserved heterogeneity from inefficiency, this method cannot be regarded as appropriate for our analysis. The cost frontier function may be further utilised to predict costs and decompose total factor productivity growth in several components, which are all interesting and relevant extensions to the main analysis. Several parametric frontier benchmarking methods will be employed to estimate the cost frontier function on a panel data set of Slovenian water distribution utilities. The COLS method, which estimates the deterministic frontier cost function, will be used to cross-check the SFA results, the latter being the main methodology used in the analysis. Hence, another important issue to study is the consistency of inefficiency scores obtained by the different methods employed.

1.4 Outline

Chapter 2 investigates the current state, major changes and problems of the water sector in Slovenia. The legal and institutional framework of the water industry is provided, while past price movements and the performance of water distribution utilities are examined. The current regulatory framework and price-setting rules in the Slovenian water industry are analysed. The relevant EU legislation relating to the water policy and water pricing is also

[5] A good review of different stochastic frontier methods is provided in Kumbhakar and Lovell (2000).

presented along with some best-practice examples of water price regulation in EU countries. Chapter 2 also provides the rationale for the regulation of network industries, examines different regulatory schemes and discusses their advantages and shortcomings. The shortcomings have to be taken into account when choosing the appropriate regulatory scheme. The focus is on incentive-based price regulation since it is shown to provide strong incentives to reduce costs and improve efficiency. Different approaches to incentive-based regulation are reviewed, namely the price cap (RPI-X regulation), revenue cap, yardstick regulation and (other) benchmarking methods. It is also shown how the regulator can overcome informational asymmetry problems associated with price-cap regulation by employing yardstick competition or some form of benchmarking analysis. Thus, a close link between regulation and efficiency is established.

Chapter 3 explores different concepts of (in)efficiency, namely technical inefficiency, allocative inefficiency, cost inefficiency and scale inefficiency. Closely related to the concept of technical efficiency, the distance function is introduced. Similarly, the cost frontier is defined, which is a standard against which to measure cost efficiency. Cost-minimisation behaviour and the properties of the cost function are studied in more detail. The use of short-run vs. long-run cost functions is discussed. Further, the optimal size of the firm and associated economies of scale and scale efficiency are examined. Cost subadditivity, which is a proper criterion for a natural monopoly, is also discussed. Finally, the distinction between economies of output density, economies of customer density and economies of scale (or, more accurately, economies of size) is made, this being particularly important in the case of network industries.

Chapter 4 studies different functional forms for estimating the cost function. Traditional and flexible functional forms are reviewed and the criteria for choosing the most appropriate functional form are provided.

Chapter 5 examines parametric frontier benchmarking methods for measuring cost (in)efficiency. The deterministic COLS method and several SFA methods are considered. Depending on the nature of the data set stochastic frontier methods are classified in cross-sectional and panel data models. Special stress will be put on panel data models which allow for many different possibilities for estimating the cost frontier and corresponding inefficiencies. The panel data models are shown to be able to distinguish between inefficiency and unobserved heterogeneity captured by firm-specific effects, but not all panel data models do this equally well.

Chapter 6 estimates the frontier cost function of Slovenian water distribution utilities. Several parametric frontier models are employed. A literature review of studies analysing the costs of water distribution companies is used as a starting point for specifying the total cost function of Slovenian water distribution utilities. After the choice of the preferable functional form is made, the cost frontier function is estimated by several different methods.

Chapter 7 continues the analysis from the previous chapter. Based on the cost frontier estimated by several models, the cost inefficiency scores of utilities are obtained and their consistency analysed. The choice of the best performing method is made. The presence of economies of output density, customer density and economies of scale is also examined. In addition, alternative uses of the cost function are explored, namely for the purposes of cost predictions and the total factor productivity growth decomposition. It is shown how the cost frontier function can be employed to decompose TFP growth into several components, namely cost efficiency change, technical change and scale efficiency change. We as well evaluate results from the economic policy-making perspective and discuss how the results could be applied in the price-regulation process.

Chapter 8 summarises the main findings of the thesis and provides final conclusions on the cost efficiency and price regulation of Slovenian water distribution utilities.

2 Organisation and Regulation of the Slovenian Water Industry

In this chapter we present the Slovenian water industry, which is part of the Slovenian communal sector providing public services of drinking water supply, wastewater treatment, solid waste management and some other services. We begin by describing the legal and institutional framework of the water industry and explore the current state, performance, price movements, major changes and problems of the communal sector in Slovenia. Before we investigate the price regulation of Slovenian water distribution utilities, different regulatory schemes are briefly reviewed – the traditional rate of return regulation and incentive-based regulation. Since the main focus of the thesis is efficiency, one of the central issues to be examined is the link between regulation and efficiency. Thus, the efficiency analysis introduced in the later chapters is put into the context of regulation. Yardstick competition and benchmarking is shown to be a useful tool in order to obtain the efficiency estimates needed for (incentive-based) price regulation. Since on 1 May 2004 Slovenia became a new EU member state it also has to comply with the EU legislation. We thus proceed by describing the relevant EU legislation concerning water-pricing issues. In addition, we provide some best-practice examples of water price regulation in EU countries. Finally, the current regulatory framework and price-setting rules in Slovenia's water industry are analysed.

2.1 Current State and Organisation of the Slovenian Water Sector

2.1.1 Institutional Framework

In Slovenia communal services, i.e. services related to water supply, wastewater treatment and solid waste collection and disposal, are generally managed at the local community level. They are classified as obligatory local public services since municipalities or local communities are obliged to provide these services. The core legal, regulatory and institutional framework addressing the issue of communal services and communal service providers in Slovenia consists of the following laws: the Public Utilities Act (adopted in 1993), the Environment Protection Act (1993, 2004), the Law on Local Self-Government (1993), the Law on Financing Municipalities (1994), and the Law on Prices (1991) that was

replaced by the Law on Price Control (1999). The water sector is further regulated by the Water Act (2002, amended in 2004), which replaced the old Water Act (1981).

Communal services in Slovenia are generally provided by public utilities. They can also be delegated to private entities at the local level, with the local community remaining responsible for regulating the service providers. The local community must control the quality of services and the prices charged.

Public utilities have the exclusive right to provide public services in the territory of one or more local communities, which makes them local monopolists. Public utilities finance their operations in one of the following ways:

- by charging prices for the public services delivered; the charge can also take the form of a tariff, tax, indemnity, compensation or reimbursement;
- by receiving funds from the national budget or local community budget; and
- from other sources.

Prices are charged for the use of public services which are measurable and for which users can be defined. Public utilities can differentiate prices across different classes of consumers and with respect to the quantity used/supplied and the regularity/frequency of use. Prices can be subsidised provided that the amount and source of the subsidy are defined. Usually, public utilities receive subsidies from the national or local community's budget. Public services which are not measurable and for which users can not be defined are financed from the budget. The infrastructure of public utilities can also be financed through short-term or long-term borrowing, where a loan is taken out by the state or local community.

With respect to the ownership of communal sector public utilities, local communities are majority equity holders in most public utilities, while in some utilities they hold a minority share. The far most frequently utilised legal form of public utility that provides public services, in particular communal services, is the public enterprise.[6] It is followed by concessions and joint ventures involving public capital. Usually, in smaller municipalities all communal activities are joined within a single company while in larger municipalities communal activities are provided separately by several companies.

[6] The capital city of Ljubljana is the only one that provides local public services through a public holding company which consists of seven public enterprises providing the following public services: gas and heating, drinking and sewerage water, collecting and disposal of solid waste, city's public transport, cemetery, and marketplaces.

2.1.2 *Major Changes and Unresolved Problems Affecting the Sector's Performance*

Before 1991, 'socially-owned' public enterprises were responsible for providing communal services as well as for investing in communal infrastructure.[7] In many respects these enterprises were operating in a very similar way to any other enterprises of the Slovenian economy. Although the prices of communal services were controlled by the state at that time, they enabled enterprises to cover their operating costs and, besides that, the sector had two stable sources for funding communal infrastructure investments: depreciation charges and transfers from a state fund established especially for this purpose. The post-1991 period has witnessed dramatic changes in the legal environment and institutional framework in which local communities and their respective communal sector public utilities operate. These changes can be classified in two man groups: changes in the relationship between local communities and communal sector public utilities and changes in local communities' legislation. Both had unfavourable impacts on the level of communal infrastructure investments (Mrak, 1997).

Under the 1993 Public Utilities Act, the ownership of communal infrastructure was transferred to local communities meaning that they became owners of assets previously in the possession of communal sector public utilities. Accordingly, local communities as the new owners of communal infrastructure have become responsible for the investment required to maintain and upgrade communal infrastructure. Although at the conceptual level the 1993 law clearly defines the ownership rights of local communities regarding communal infrastructure, the related bylaws have not provided any precise guidelines for the law's effective implementation. There were, for example, no clear accountancy instructions for the depreciation of communal infrastructure. As a result, depreciation has often not been properly included or has not been included at all in the financial statements of the new owners of infrastructure. The solution came in 1997 with the inclusion of special accounting

[7] Before 1991, Slovenia was part of the Socialistic Federative Republic of Yugoslavia and had a market-planned economy. One of its peculiarities was that enterprises were owned neither by the state nor by individuals but by society as a whole. It was often said that enterprises were owned by everybody in general and nobody in particular. Thus, the enterprises did not have proper owners. After the disintegration of Yugoslavia in 1991, Slovenia became an independent state and entered into the transition period. The transition from a market-planned to a market economy was among others followed by the transformation of ownership, i.e. by nationalisation or the privatisation of previously socially-owned public enterprises.

standard (SRS 35) to the Slovenian Accounting Standards to deal with issues specific to the public sector. According to the SRS 35, utilities can either rent or manage the infrastructure. In the first case, public utilities have to pay the rent and the infrastructure does not enter their balance sheets while, in the second case, the infrastructure enters their balance sheets and they are responsible for depreciation.

Another problem negatively influencing investment in communal infrastructure in Slovenia in recent years is the local community reform. As an integral part of this process, local communities were given new responsibilities yet the funding was not increased to account for these new duties. Further, under the 1993 Law on Local Self-Government the large majority of local communities were divided into two or more local communities. The number of local communities grew from 62 in 1991 to 192 in 2003. Since the resulting local communities are often very small in terms of their population (more than one-third of local communities in Slovenia have a population of less than 5,000 inhabitants) and weak in terms of their economic base, many do not have sufficient financial resources to invest in communal infrastructure. The issue of infrastructural investment is further complicated by the fact that the distribution of assets and liabilities of old local communities among the new ones has not yet been fully completed in all but a few cases. Some public utilities face the situation where they operate in more than one local community which exposes them to various problems as they have to deal with more than one owner of communal infrastructure and often with different legal regimes and tariff systems.

Besides the problems related to the investment process, the poor financial performance of communal sector public utilities is another issue. Price determination, which is under government control, was based on political considerations and other macroeconomic goals (e.g. reducing inflation) so the prices of communal services in the post-independence period were increasing slower than the inflation rate, and public utilities providing communal services were not allowed to increase their prices to the full-cost level (Štruc, 1997). Consequently, the financial health of communal sector public utilities was, and still is, well below the financial performance of the Slovenian economy as a whole. In fact, most communal sector public utilities are unable to cover the total costs they incur and therefore operate at a loss, which is highly unsustainable in the long run. Some relevant indicators of the size, performance and prices of the Slovenian water sector are given in Appendix I.

2.1.3 Water Prices in the Post-1991 Period

Since early 1992, the prices of communal services have been continuously under the control of the Ministry of Economic Relations and Development. Several decrees, rules and guidelines on the price determination of communal services have been issued by the government. Although various price adjustment procedures have been applied over the whole period up to the present time, their common feature has been to allow the prices of public utility services to rise by a rate lower than the industrial producer price index for the respective period (i.e., in real terms the prices have been decreasing). If the prices of communal services were not sufficiently high to cover current expenditures, utilities were, as an exception, allowed higher price increase. The prices of water supplied over the 1991-2000 period are provided in Appendix I (Table I.1).

This price-setting policy has not allowed communal service providers to cover all their operating and capital costs. In most cases, tariffs have only been sufficient to cover current expenditures but not to finance regular maintenance and the replacement of fixed assets, not to mention new investments (Hrovatin, 2002). In an analysis carried out by Kavčič (2000) it was discovered that in 1998 the average costs were on average 30 percent higher than the average price of water supplied to different customer groups. Moreover, most public utilities providing drinking water operated at a loss. Similar findings were also reported for wastewater treatment, as well as the collection and disposal of solid waste.

In order to improve their poor financial position caused by the restrictive price regulation, communal service providers had to find 'creative' solutions to find new funding sources. One of the most commonly used practices over the examined period was the introduction of local environmental fees aimed at tackling the specific pollution concerns of the local population. Another issue to be pointed out is that there are some striking differences between the prices charged in different local communities. This arises from the fact that public utilities had very different starting positions before price control was introduced (see Appendix I, Table I.1). However, this issue failed to be recognised by the relevant authorities. So, after the price control was put in place these utilities were put in an unequal position and were faced with different operating environments (Hrovatin, 2002).

As in many transition economies, another distinguishing feature of the prices for communal services in Slovenia is that there are significant differences in price levels for water supply between different customer groups. The prices for water supply have been by far the lowest for households and the highest for businesses. However, the range of prices paid by different

customer groups has narrowed over time (Štruc, 1997). Different prices for different groups of customers may sometimes be justified by differences in the costs of providing services. However, the policy of subsidising households and thereby addressing social policy issues seems to be a more plausible explanation of why the existing tariff structure weighs more heavily on businesses than on households. This should, however, be weighted against the distorting effects of low-priced services. Not only do customers receive misleading signals as to the real value of services, reducing their incentives for efficiency, but subsidisation also results in the deferment of urgent investment in communal infrastructure while private capital is not interested in entering such sectors and participating in the investment process.

Clearly, the strategic objective of the communal sector in Slovenia should be to move towards the reliable and cost-efficient provision of communal services which will duly take account of the security of supply, safety of the population and protection of the environment and be in compliance with the relevant EU legislation. In order to achieve this objective, a whole range of co-ordinated policy measures has to be designed and put into operation. The following core elements of the communal sector's transformation consistent with the EU's legislation and regulatory framework have been identified (Mrak, 2000):

- introduction of cost-reflective prices of communal infrastructure services and tariff reform;
- introduction of competition for the market, restructuring of service providers and private sector involvement;
- a legal and institutional framework which would provide clear rules for private sector involvement in communal infrastructure investment and in the provision of communal services; and
- a regulatory framework: independent yet accountable regulatory authorities are needed to oversee utilities (instead of supervision by ministries).

In what follows we primarily focus on issues related to the regulatory framework and price setting methodology in the Slovenian water sector. Before we analyse this issue, we briefly discuss the different regulatory schemes that have been the most commonly used in the water sector. Also, the EU legislation relating to water pricing will be presented together with two best-practice examples of price regulation in the EU water sector. These schemes will then serve as benchmarks against which the current Slovenian price regulation design will be judged.

2.2 Different Price Regulation Schemes

Network industries demonstrate substantial economies of scale since by their very nature they are extremely capital-intensive. As fixed costs form a substantial amount of total costs, it is cheaper for a single company to provide services associated with the network. Network industries thus provide a clear example of natural monopolies.[8] The competition in this case is ineffective since new entries would lead to the problem of excessive entry that involves the needless duplication of fixed costs associated with the network (Baldwin and Cave, 1999). Even if the existence of the natural monopoly is justified, economists have long recognised that a monopoly does not lead to the desired market outcome since it results in production and allocative inefficiencies. The monopolist will, if left alone, set higher prices and sell lower quantities of output compared to the situation in a competitive market. The recognition of these problems, among other issues, has provided sound justification for the need for the price regulation of network industries in order to prevent the abuse of monopoly power. This has led to a long history of attempts to regulate natural monopolies and to a vast literature discussing the problems and attempts at regulation (Netz, 1999).

In practice, there are two main forms of regulatory control: the traditional rate-of-return (RoR) regulation or cost-of-service (CoS) regulation, and incentive-based or performance-based regulation (e.g., price cap, revenue cap). The rate-of-return regulation originates from the US where it has a long tradition. Utilities there tend to be privately owned rather than embedded in government departments and have been subject to regulation by specific rate commissions. The existing regulatory structure has been established over a long period of a formal legal procedure dating back to the late 19th century (Crew and Kleindorfer, 1986). Historically, most other countries chose public ownership as the common method for controlling natural monopolies. Because of recent shifts towards privatisation worldwide, other countries have also had to establish regulatory agencies and develop a regulatory methodology to oversee the newly privatised firms. As a response to some serious flaws in the rate-of-return regulation, alternative regulatory schemes such as price-cap schemes have been taken into consideration. Although a few American precursors can be identified, price-cap regulation (RPI-X regulation) was first applied on the large scale in the UK. Price capping was initially developed as a temporary control mechanism in the transition to full competition for UK telecommunications (Littlechild, 1983) and then extended to other UK

[8] The formal definition of a natural monopoly and economies of scale will be given in Chapter 3, Section 3.5.

utilities as they were privatised.[9] Yet price-cap regulation is also not without its shortcomings. Further, since regulation and performance are closely related one of the key issues is how to obtain reliable efficiency and productivity indicators to be entered into the price-cap formula. Incentive-based regulation is thus frequently combined with some form of benchmarking analysis. In the water industry, for example, the rate-of-return regulation is typically applied in the US (Mann, 1993), the UK exercises price-cap regulation combined with benchmarking (OFWAT, 1999) while in Italy benchmarking or yardstick competition is used in the price-setting process (Massarutto, 1999).

2.2.1 Rate-of-Return Regulation

Rate-of-return regulation aims at preventing the exploitation of consumers by a company with monopoly power. It tries to prevent the abuse of monopoly power, that is the ability to earn excess profits by setting prices above the current long-run cost of supply for an efficient company. Accordingly, the regulator sets prices for the utility in such a way that they cover the utility's costs and include a rate of return on capital that is sufficient to maintain the investor's willingness to replace or expand the company's assets. This is why RoR regulation is also known as cost-plus regulation (Baldwin and Cave, 1999). While this form of regulation is apparently simple and seems as if it would achieve feasible average-cost prices, it will be discussed below that in practice average-cost prices are not achieved because the regulatory mechanism in fact gives firms an incentive not to minimise their costs.

Regulators determine the revenue required to give the utility a fair rate of return and then set prices so as to recover this revenue. This can be stated by means of the following cost-plus formula (Hill, 1995):

$$Rev = OPEX + Dep + T + ROR \times (RB - AD) \tag{2.1}$$

where:
Rev – required revenues
OPEX – operating expenditures
Dep – depreciation expenses for the current year
T – taxes
ROR – allowed rate of return

[9] Armstrong, Cowan and Vickers (1994) described incentive regulation in the UK.

RB – rate base, i.e. gross value of the utility's property (plant investment, including an allowance for working capital)
AD – accumulated depreciation

The calculation in Eq.(2.1) can be based either on historical accounting costs or projected costs. The value of the rate asset base depends on whether the regulator uses original costs, which is the typical approach, or replacement costs. If the regulator does not consider an item of plant used and useful, it may not be allowed in the rate base. A utility is allowed to earn revenues to cover all expenses plus a return on its investment in useful plant and equipment. Fixing 'fair' or 'reasonable' rates of return involves a balancing of the investor's and consumers' interests. For instance, from the point of view of an investor of company it is important that there be enough revenue not only for operating expenses but also for the capital costs of the business. The owner should be entitled to the return on equity that corresponds to the return on investments in other enterprises with comparable risks (Crew and Kleindorfer, 1986).

Once an estimate of the overall level of required revenues is obtained, total operating and capital costs are assigned to different service classes (e.g. residential and business customers). This is done on the basis of cost-causation principles. Prices for each class of a utility's service *i* (P_i) are the ratios of allocated costs (C_i) to historical or estimated sales (S_i):

$$P_i = C_i / S_i \tag{2.2}$$

where sales are expressed in physical units sold to a given class of final customers. Because the assigned costs are typically accounting costs rather than economic costs there is usually some arbitrariness in assigning them to individual customer groups (Hill, 1995).

In summarising, the general principle underlying RoR regulation is that the firm should recover its costs (i.e., $P = AC$) and consumers should pay a fair price, with fairness argued on the basis of the cost-causation principle. The emphasis on fair rates seems to be much more closely related to equity than to economic efficiency. Neither efficient pricing nor productive efficiency seem to be directly addressed by RoR regulation. Since RoR regulation does not directly provide incentives for efficient production, at the least the regulator can attempt to control the size of all variables on the right-hand side of Eq.(2.1), disallowing certain items if they are excessive. While regulation does not seem to be very concerned with promoting productive efficiency in either a traditional static sense or in a dynamic sense by providing incentives for research and development, it does, however, achieve a stable basis for

operations as the companies are prepared to make the investments required (Crew and Kleindorfer, 1986).

Rate-of-return regulation is subject to several flaws. The first, obvious disadvantage has already been established. There is a lack of an incentive to reduce costs and operate efficiently since the company knows it will be able to recover any growing costs through a subsequent price increase. Prices are primarily based on historical costs. Provided that price reviews take place often enough the firm pays no penalty for inefficiency. On the other hand, the firm benefits little from any efficiency gain. This arises because, if any cost savings are made, they will almost immediately be taken from the firm and given to consumers in the form of lower prices (Baldwin and Cave, 1999).

The second disadvantage is the bias towards capital-intensive production methods. This phenomenon is also known as the Averch-Johnson effect. Averch and Johnson (1962) demonstrated that the marginal rate of technical substitution of capital for labour is lower for a profit-maximising regulated firm than for a cost minimiser, the latter equating this quantity to the cost of capital to the wage rate ratio. It follows that under this regulation capital is over-utilised and labour is under-utilised relative to any cost-minimising solution. Regulated firms have an incentive to over-invest in capital equipment since this expands the 'rate base' (i.e. the value of capital employed) against which the rate of return is measured and so allows higher absolute levels of profit for a given relative rate of return. As a result, this will skew inefficiency in the direction of the excessive use of capital (Baldwin and Cave, 1999).

The Averch-Johnson result is important since it shows how rate-of-return regulation introduces inefficiency to the capital-labour ratio, i.e. allocative inefficiency. If the Averch-Johnson effect were, *ceteris paribus*, the only result of RoR regulation, the regulation would be an obvious failure. However, regulation can increase output and reduce prices sufficiently so that, even though costs are not minimised, there is a welfare gain from regulation. It is argued that the primary concern of RoR regulation is with the equity aspects of monopoly, preventing the monopolist from exploiting consumers. It is also not possible to say in general that the output or capital intensity of the regulated firm will exceed that of a pure monopolist (Crew and Kleindorfer, 1986).

Another disadvantage of RoR regulation is the extremely detailed nature of regulation in terms of defining the rate base and monitoring the rates of return actually achieved. The regulatory intervention is the commission process dictated at determining a fair rate of return, together with a definition of the rate base and allowable expenses. There is a scope for a substantial amount of 'gaming' between the regulatory commission and the utility. Rate

hearings occur on a frequent basis necessitating a large amount of effort on the part of the regulated utility and the regulatory body.

2.2.2 *Price-Cap Regulation*

Price-cap or RPI–X regulation has emerged as an alternative to rate-of-return regulation. In large part the problems arising in the RoR regulation are due to the fact that regulated firms do not have an incentive to operate at minimum cost. Price-cap regulatory schemes attempt to focus regulatory scrutiny on efficiency and performance improvements rather than the rate of return on the rate base, as in RoR regulation. While Littlechild (1983) developed a practical basis for RPI-X regulation, Vogelsang and Finsinger (1979) developed the theoretical basis for price-cap regulation.[10] Essentially, the advantages of price control are that it avoids the disadvantages of rate-of-return regulation. Under price-capping, prices are set in advance for a period of three to five years allowing the firm to benefit from any cost savings made during that period, but they are then recalculated at regular intervals in order to bring them back into line with the underlying costs. The price cap usually permits a utility to increase its overall level of prices by the previous year's rate of inflation, as measured by the retail price index (RPI), minus a percentage productivity factor (X) that reflects the real-cost reduction the regulator expects. Prices (P_t) are therefore set according to the following formula:

$$P_t = (1 + RPI - X) \times P_{t-1}. \tag{2.3}$$

Unlike RoR regulation, price-cap regulation does not simply allow a firm to recover whatever cost it has historically incurred. Instead, the regulator makes cost projections into the future and sets overall prices so that they recover those expected costs. Price-cap regulation is thus forward-looking. If the regulated firm is able to increase its efficiency and reduce costs more than the regulator anticipates, its profits will go up. If it is less efficient than expected, its profits will go down. This system of 'periodic' price capping has been shown to give strong incentives to improve efficiency, thereby earning economic profits. Through economic profits, improved efficiency is revealed to the regulators who take this information into account at the next periodic review of price limits. The benefits arising from the company's lower costs due to improved efficiency can thereby be passed on to

[10] They developed a regulatory mechanism similar to the price cap. For discussions of price-cap regulation, see Vickers and Yarrow (1988).

consumers in the form of lower prices. The *quid pro quo* of higher profits today is lower consumer prices tomorrow (Vass, 2000).

Price-cap regulation refers to a class or type of rate regulation, not to a specific scheme. The precise nature of this type of regulatory scheme depends upon how the basic elements in the rate-cap formula are defined. Some schemes apply the cap to total revenue (P in Eq.(2.3) is replaced by $P \times Q$), while other schemes apply the cap to P, the weighted average price. In either case, the firm can increase profits by reducing costs. In the latter case, it can also increase profits by increasing the quantity sold. A scheme for price caps is simpler than for revenue caps and avoids the need to forecast volumes. It also provides an incentive to the regulated firm to serve new customers and develop new business. On the other hand, a price cap shifts the risk of fluctuations in system usage to the regulated firm. A revenue cap mitigates this risk (Hall, 2000).

Price-cap rates are set at fixed, periodic intervals, typically three to five years. This practice contrasts with RoR regulation where the period between rate cases is usually variable and in general controlled by the utility. Frequent reviews will tend to undermine the incentive properties of price-cap regulation, while infrequent reviews create the possibility of prices deviating from costs over an extended period of time. The initial revision of the price structure and price levels prior to the beginning of price- cap regulation is essential. A regulatory failure may be at least partly avoided by setting initial prices to reflect the costs prior to imposing the RPI–X regulation. Under price-cap regulation the regulated prices are set at a level that enables the generation of sufficient revenues to cover all justified costs of the utilities.

The productivity improvement in the price-cap formula, the X factor, assures that productivity improvements are passed on and that existing above-normal profits and cost inefficiencies are removed. The X factor assures that customers receive some price benefits as a result of price-cap regulation and that management will have to achieve some target level of efficiency improvement before stakeholders benefit from enhanced profits as a result of lower costs and/or additional sales. Usually, the X is set to reflect the expected growth in total factor productivity based on past TFP growth. The greater the X, the tighter is the constraint. The individual X-factor is usually based on two pieces of information – on the rate of productivity growth reported in the industry in recent years and on the firm's cost inefficiency, i.e. on the extent that a given firm is operating below the best practice in the industry (Coelli et al., 2003). Productivity growth is therefore a broader concept than efficiency improvement. In some cases, regulators only consider the latter component of TFP growth.

However, since only imperfect information is available to the regulators they can merely observe the firm's actual level of costs while the firm itself has an accurate view of what it can achieve. Regulators can remedy this problem by obtaining better information about the firm's productive potential. In setting the yearly efficiency factors, the regulator could perform some form of benchmarking analysis of the utilities' costs.[11]

There are different ways to translate the efficiency scores, θ, obtained by the benchmarking analysis into the X-factors. An extreme case would be to impose a direct link. In the case of cost benchmarking, the efficiency score θ reveals a utility's efficient level of costs as opposed to its actually incurred costs. In other words, the utility could on average produce the same level of output at $\theta \times 100\%$ of its current costs. The X-factor could then reflect the path from the actual cost level towards the efficient cost level through a gradual decrease in prices through time. If T is the duration of the regulatory period in years and θ the efficiency score obtained by one of the benchmarking techniques, then the yearly X-factor would be set such that:

$$(1-X)^T = \theta . \tag{2.4}$$

Firms that operate at higher productivity levels would be given a higher efficiency score and consequently a lower X-factor reflecting the expected reduction in costs as a result of an increase in productivity. This introduces a degree of competitive pressure. The link between the benchmarking results and the X-factors can also be less direct. If the regulator believes it can only imperfectly perform a benchmarking analysis, rather than directly deriving the X-factors from corresponding efficiency scores, it may wish to use the results merely as a starting point for setting the X-factors. The benchmarking results would in such a case provide information on the range in which the X-factor could be located.

The regulator can also explicitly take into consideration the network quality and reliability of supply requirements. This is because profits can be increased not only by reducing costs but also by reducing service quality. With RoR regulation there is an incentive to 'gold-plate' service quality and reliability. Thus, defining a minimum standard is not a problem. With price-cap schemes it is vital that regulators define minimum standards and enforce them in order to prevent the temptation to 'cut corners' (Hall, 2000).

[11] Recall Section 1.3, where different benchmarking methods are presented.

While the price cap mechanism gives firms some incentive for efficient production, it is not, however, without its own problems. If firms recognise that prices ultimately follow costs, they may well not reduce costs to efficient levels. For this reason, a regulator needs reliable estimates of the productivity or efficiency potential of the regulated firm. This can be obtained by employing yardstick competition or some form of benchmarking analysis, as will be further discussed in the next section. From the earlier discussion it can also be concluded that price-cap regulation does not exempt the regulator from the need to examine the utilities' cost of capital and whether the allowed revenues will provide an adequate return on capital. The price-cap regulation is in this sense similar to RoR regulation since both models require the regulator to address the same issue, that is the need for the regulatory scheme to ensure that the approved rates allow investors to recover the cost of the capital they provide. It should also be noted that RoR regulation is not an automatic cost-pass-through process. Rather, fixed rates are established and apply until they are changed. Usually the utility proposes the changes and the changes are approved by the regulator, often with modifications. Thus, because of the 'rate lag' there is another similarity between rate-of-return regulation and price-cap regulation. Nonetheless, price-cap regulation formalises and simplifies the rate review process and focuses regulation on efficiency improvements in a way that differs from RoR regulation. The main advantages of price-cap regulation are that it provides incentives for efficiency improvements, provides for fixed periods between adjustments and can be simpler to implement than RoR regulation. On the other hand, the disadvantages are that it requires the regulator to establish and enforce performance, quality and reliability standards and it does not avoid the regulator having to ensure that the allowed prices will yield an adequate reward for the investors.

In the literature, there is also some theoretical evidence on the performance of price cap regulation. Based on several developed models on the optimal regulatory policy it can be concluded that price-cap regulation can give firms a larger incentive to invest in cost reduction than rate-of-return regulation and there is some evidence that price-cap regulation is or can be part of the optimal regulatory mechanism.[12] However, while price-cap regulation does have the benefit of increasing the incentive for firms to invest in cost reduction, it induces a variety of new problems. Subsequent theoretical work and experience from the implementation of price-cap regulation suggest there are also some problems with this scheme arising from the proper calculation of the price index and the optimal pricing structure, the impact on quality, and renegotiations and end-game problems. For example, the end-game problem may arise when a firm is regulated under a price-cap regulation for

[12] See, for example, Cabral and Riordan (1989), Sibley (1989), Lewis and Sappington (1989), and Clemenz (1991).

some known, finite period of time, after which rate-of-return regulation is implemented. One problem here is that the firm may manipulate the system by shifting costs into the future (Netz, 1999).

Part of the reason price-cap regulation works is by allowing the firm to keep the gains it makes from cost reductions. If regulated firms believe that price caps will be revised to appropriate the gain from cost savings, then their incentive to invest in cost savings is reduced. The length of the regulatory period also plays a role. In fact, there is a trade off: the longer the amount of time between reviews, the more likely it is the price will be low at the next review (as the firm has a greater pay off in investing in cost reduction since it retains profits for a longer period), but the amount of time during which the price is high is longer. Similar conclusions were made in Laffont and Tirole (1993) where the extensive use of economic modelling is used to derive optimal incentive mechanisms. The regulatory problems are analysed within the principal-agent framework which provides a valuable insight into the role of information as a source of monopoly rents. One of the main conclusions from their analysis is that there is a basic trade-off between incentives and rent extraction. Being unable to monitor the firm's effort and having less information than the firm about its technology, the regulator has to promote cost reduction and extract the firm's rent. As the firm improves its efficiency it earns more profits, but the incentive to improve efficiency would not be there if it were not able to keep these profits in the first place. Hence, the higher the rents that would stay with the firm the greater the incentive to improve efficiency. On one hand, a powerful incentive scheme (e.g. fixed-price) provides strong incentives to increase efforts, but only the firm has the benefit from this in terms of higher profits while customers do not see any gains. Under a weak incentive scheme (e.g. cost-plus), there are no incentives to improve efforts since there is full rent extraction. Thus, a trade-off between incentives and rent extraction is present.

2.2.3 Yardstick Competition and Benchmarking

With perfect information the regulator could simply mandate efficient behaviour to the firms. However, given more realistic assumptions on the costs of obtaining information regulators will necessarily be considerably less informed than firms on such matters as technology, cost and demand conditions. Thus, a principal-agent problem under the incomplete information arises, whereby the regulators must use procedures which employ data that is relatively easily obtainable (e.g. arising from audited accounting records) to define incentives for

efficient behaviour. One way to obtain the information is to employ yardstick competition or some (other) form of benchmarking.[13]

Yardstick competition is a regulatory instrument that can be used if direct competition between agents does not exist or does not lead to desirable outcomes. More specifically, it can be used where firms have little incentive to promote cost efficiency. An important reason to use yardstick competition is the existence of market power due to regional monopolies; typical examples are network industries. The framework used in the analysis is the principal-agent model. The regulator rewards agents on the basis of their relative performance and therefore generates incentives for promoting efficiency. Agents are forced to compete with a 'shadow firm' whose performance is determined by the average practice in the industry. If the regulated firm is on average more efficient than the firms to which it is compared, it will make above-normal profits.

The theoretical framework was developed by Shleifer (1985) who proposed yardstick competition to regulate local or regional monopolies that produce a homogeneous output. Following Shleifer (1985), a one-period model is considered, with N identical risk-neutral firms operating in an environment without uncertainty. Each firm faces a downward-sloping demand curve $q(p)$ in a separate market. Each firm has an initial constant marginal cost c_0 and can reduce it to constant marginal cost c by spending $R(c)$. It is assumed that $R(c_0) = 0$, $R'(c) < 0$, and $R''(c) > 0$. Thus, the higher is the investment in cost reduction the lower is the final unit cost, where cost reduction is cheap at the start but gets progressively costlier. As reduction expenditures result in fixed-cost, firms have decreasing average costs. The profits of a firm are given by:

$$\pi = (p - c)\, q(p) + T - R(c), \tag{2.5}$$

where T is a lump-sum transfer to the firm. If the regulator chooses c, p and T so as to maximise the sum of consumers' surplus and the firm's profits, then the solution is the social optimum given by:

$$\begin{aligned} &R(c^*) = T^*, \\ &p^* = c^*, \\ &-R'(c^*) = q(p^*). \end{aligned} \tag{2.6}$$

[13] Yardstick competition is also known as competition by comparison, and benchmarking can be referred to as a comparative analysis.

Thus, the transfer just covers the expenditures on cost reduction while prices are set to equal the marginal cost. The last equation says that the marginal cost of reduction has to be equal to the output. Intuitively, lowering unit costs by Δc and thus reducing production costs by $q(p)\Delta c$, requires $-R'(c)\Delta c$ investment in cost reduction. At the optimum, the costs and benefits of marginal change in c must be equal.

To order firms to achieve c^*, the regulator must know $R(c)$, which is seldom the case. On the assumption that the regulator does not have this information, it is shown that rate-of-return regulation fails to deliver any cost reductions at all. Under rate-of-return regulation, the regulator sets $p = c$ and $T = R(c)$ whatever the costs are. Faced with this policy, managers recognise that their profits are going to be zero regardless of their costs and, since they prefer not to reduce costs (i.e., they prefer to minimise their effort given the level of profit), they keep $c = c_0$.

As an alternative to rate-of-return regulation, Shleifer (1985) introduced yardstick competition. He proposed to eliminate the dependence of the firm's price on its own chosen cost level by using cost levels of identical firms to determine the price. Each firm i is assigned its own shadow firm, with the cost level being equal to the mean marginal cost of all other firms, $\overline{c}_i$, and with a similarly defined cost-reduction expenditure, $\overline{R}_i$. This shadow firm serves as a benchmark in the yardstick competition. Yardstick competition thus seeks to provide an incentive for utilities to strive for lower costs by inducing them to compete with one another for cost reductions. Shleifer (1985) showed that if the regulator sets prices so that $p_i = \overline{c}_i$ and uses transfer rule $T_i = \overline{R}_i$, then the social optimum is achieved as the unique equilibrium of a game in which firms simultaneously choose their unit-cost levels. The unique Nash equilibrium for each firm i is to pick $c_i = c^*$. Therefore, in the case of homogeneous output yardstick competition delivers the first best.[14]

Shleifer (1985) also demonstrated how the yardstick competition concept can be applied to firms producing heterogeneous outputs if these outputs only differ in observable characteristics (ψ). For example, in the case of network industries the heterogeneity of output consists mainly of the different characteristics of distribution service areas (e.g. network size, differences in the mix of residential and business customers, population density, and

[14] It is also shown that if lump-sum transfers are not available to the regulator, he must compensate the firm for cost-reducing expenditures by allowing higher prices. In the case of homogeneous goods this amounts to the average cost pricing version of yardstick competition in which all firms pick the second-best unit cost levels.

landscape). To correct the yardstick competition for heterogeneity, the regulator can use a multivariate estimation of the average cost function $\hat{c}_i = a + b\psi$. The observable characteristics are included as explanatory variables and will thus correct for cost differences that are only due to the heterogeneity of output. The regulator then sets corrected yardstick prices for individual firms that incorporate their differences according to the rule: $p_i = \hat{c}_i$.

In Section 1.3 we already introduced and briefly described different benchmarking methods. According to the classification of benchmarking methods provided in Section 1.3, the model of yardstick competition proposed by Shleifer (1985) makes use of average parametric benchmarking methods.[15] Besides average benchmarking, frontier benchmarking methods are also often used in price regulation. The latter have a stronger emphasis on the efficiency of the regulated firms, which is one of the reasons parametric frontier benchmarking methods were chosen as the main method in our analysis. These methods will be analysed in detail in Chapter 5.

However, the application of yardstick competition or other benchmarking methods in the price-regulation process is not without some concerns. From a regulatory point of view, it is encouraging that different benchmarking models provide the same results with respect to the utilities' efficiency. For instance, if this were not the case, any one-to-one translation of efficiency scores into X-factors in price-cap regulation would be unjustified. However, the applied economic literature reveals either mixed or negative evidence on the cross-model consistency of computed efficiency scores.[16] In a number of studies it was found that benchmarking is, to some extent, influenced by the techniques chosen, model specification and variables included in the model. Bauer et al. (1998) defined a set of consistency conditions that, if met, would make the choice of a particular method trivial. The efficiency estimates should be consistent in their efficiency levels, rankings, identification of best and worst practices, consistent over time and with competitive conditions in the market, and consistent with standard non-frontier measures of performance. However, in the absence of any consensus on the most appropriate technique to use a purely pragmatic approach would entail the combination of results from different models. In this case, rather than using efficiency estimates in a mechanistic way regulators are advised to use benchmarking as one of the instruments for incentive-regulation purposes. Benchmarking can thus be viewed as an

[15] However, it should be noted that yardstick competition has an additional requirement of linking financial consequences to the benchmarking results, whereas no such requirement is made in the benchmarking case. Only comparability requirement has to be made in order to be able to perform benchmarking analysis (CPB, 2000).

[16] For example, see Bauer et al. (1998), Estache et al. (2004), and Farsi and Filippini (2004).

effective complementary regulatory instrument in price regulation and not as the regulator's main instrument for monitoring utilities' performance.

2.3 Price Regulation in the EU Water Sector

As a new member state of the EU, Slovenia also has to comply with the EU's legislation. Therefore, the relevant EU legislation relating to the water pricing must also be taken into account in the price regulation of the Slovenian water industry. The Water Pricing Communication (COM(2000) 477 final) plays an important role in the pricing policy by promoting the use of water charging that would act as an incentive for the sustainable use of water resources and to recover the costs of water services by economic sectors.[17]

2.3.1 Water Pricing Communication (COM(2000) 477 final)

In line with the Water Pricing Communication, pricing should be designed in a way to promote the more efficient and less polluting use of scarce water resources. This would, in turn, reduce the pressure on water resources and the environment and ensure available resources are efficiently allocated between water uses. As a result, water supply and treatment infrastructure could be more appropriately sized. This means providing water services and protecting the environment more cost-effectively. Efficient water pricing would additionally mobilise financial resources to ensure the financial sustainability of water infrastructure and service suppliers, and to pay for environmental protection.

It is argued that the lack of importance attributed to economic and environmental issues in designing existing water pricing policies, as opposed to more general social or development objectives, has led to the current situations of inefficient use, overexploitation and degradation of surface and groundwater resources. To play an effective role in enhancing the sustainability of water resources, water-pricing policies need to reflect different cost types (COM(2000) 477 final):

- *Financial costs* of water services, including the costs of providing and administering these services. They include all operating and maintenance costs, as well as capital costs.
- *Environmental costs*, representing the costs of damage that water uses impose on the environment and ecosystems and those who use the environment (e.g. a reduction in the

[17] More broadly, the EU legislation relating to the water policy is laid down in the Water Framework Directive (2000/60/EC) and the Drinking Water Directive (98/83/EC).

ecological quality of aquatic ecosystems or the salinisation and degradation of productive soils).

- *Resource costs*, involving the costs of foregone opportunities which other uses suffer due to the depletion of the resource beyond its natural rate of recharge or recovery (e.g. linked to the over-abstraction of groundwater).

Overall, each user should pay for the costs resulting from their use of water resources, including environmental and resource costs. Moreover, prices should be directly linked to the water quantity used or pollution produced. Pricing structures should thus include a variable element (i.e. volumetric rate, pollution rate) to ensure that prices have a clear incentive function for users to improve their water-use efficiency (i.e., to facilitate water conservation) and reduce pollution. Further, water prices should be set at a level that ensures the recovery of costs for each sector (i.e. agriculture, households, industry). It is important to ensure that the most polluting and least efficient sectors pay for their pollution and use. A significant reduction in existing pressures on water resources can be expected through a sectoral recovery of the costs of water services.

The level of integration of economic and environmental objectives into water-pricing policies differs highly among member states of the EU, within member states and between economic sectors. Overall, the full recovery of financial costs is only partly achieved. Environmental and resource costs are rarely considered in pricing policies. Although most water-price structures for domestic water supply include fixed and variable elements and have an incentive role, flat water charges independent of use or pollution are still in use. The last few years have recorded the increasing role given to pricing in the water policies of many member states. Most European countries already apply two-part tariffs to the supply of water (Hrovatin and Bailey, 2001).

Moreover, the Commission calls for a harmonised approach to pricing within the EU. The adoption of a common definition of key cost variables would facilitate the comparison between costs and prices and benchmarking for different water services, uses and countries. Harmonisation requires standardised accountancy practices and financial costs, for example in the depreciation of capital facilities and the use of replacement (rather than historic) cost when evaluating fixed assets. It also requires the adoption of common methodologies for the monetary valuation of environmental and resource costs and benefits. High information costs are often mentioned as a constraint on the development of water-pricing policies that better account for economic and environmental objectives. It is also emphasised that the harmonisation of approaches will not result in uniform prices due to the differences in costs

reflecting geographical, topographical, climate, institutional and economic factors, which vary considerably not only between but also within countries.

As a result of the natural monopoly situation of most water suppliers (whether public or private), control over the water prices charged to consumers is necessary to ensure that prices adequately reflect existing costs and do not hide inefficiency. Benchmarking that compares the quality of water services, costs and prices is another key element of a communication strategy. Benchmarking of the suppliers' performance can act as an incentive for them to improve their efficiency and quality of services and reduce their costs and prices.

Decision-making with respect to water prices varies considerably between and within EU member states. Water-price levels and structures can be decided at the local, regional or national level. In most countries, price setting is decentralised either to municipalities, ministries or independent economic and/or environmental regulators. Nevertheless, it is rare that decisions on pricing are entirely decentralised with no supervisory power institutionalised at the national (federal) level. Municipal decision-making will inevitably lead to a greater diversity of pricing practices within a given country than centralised decisions by government ministries. To the extent municipalities have an influence over price levels, local interests may predominate over the regional or national interest. Likewise, independent regulatory authorities may have different perspectives on price setting than respective ministries, the former being perhaps more professional and technical in their approach and the latter being more political and bureaucratic. In addition, separate economic and environmental regulators may differ in prioritising the interests of the environment and water utilities (Hrovatin and Bailey, 2001).

Below the two best-practice regulatory examples applied in the EU water sector are briefly presented, more specifically the UK and Italian price regulation schemes. So far, other regulatory authorities in general do not make use of incentive-based price-regulation approaches. The Water Pricing Communication will hopefully facilitate some changes with respect to this issue.

2.3.2 Price Regulation in the UK

Water companies in England and Wales were privatised in 1989 under the 1989 Water Act. The 1989 Act preserved the local monopoly status enjoyed by the water companies, providing a rationale for their regulation. The companies are closely regulated by a number of bodies to ensure drinking water quality remains high, the environment is protected and

improved, and the customer receives improved standards of service. There are 10 water and sewerage companies and 17 water only companies responsible for water supply, sewerage and sewage treatment and disposal (OFWAT, 1999). Prices of the water industry are regulated by the Office of Water Services (OFWAT) according to the RPI+K price cap. Under the RPI+K formula a regulated firm's prices are allowed to rise at the rate of inflation, plus an amount K which reflects investment needs and expected efficiency gains, and can take on positive or negative values. The K factor comprises two elements: a factor X which reflects future efficiency gains of usual utility operations, and a factor Q to allow for mandatory improvements in quality standards and the environment. Hence, the formula takes on the form of RPI–X+Q (OFWAT, 1993). The price cap for water companies applies to a weighted average set of a basket of services (the tariff basket formula). The first price cap, for the 1989-1995 period, was set by the government. Subsequent price caps were set by the OFWAT (Kennedy, 1997).

The OFWAT sets prices on the basis of yardstick competition. Price limits are set for five-year periods with all companies being expected to achieve significant efficiency gains and with tougher targets being set for more inefficient companies. If the companies outperform their efficiency targets, then the shareholders earn higher returns in the five-year period at the end of which customers benefit through further reductions in their bills. At the OFWAT's 1994 Periodic Review, an OLS analysis was carried out to estimate the total operating expenditures of the water companies (Williamson and Toft, 2001).

At the 1998 Periodic Review an OLS analysis was carried out separately on (controllable) operating costs and capital-maintenance costs. Based on the results from different models, companies are ranked in performance bands. According to Table 2.1, companies are ranked on percentage cost differences vis-à-vis the yardstick with respect to both operating and capital-maintenance costs. Individual company circumstances are taken into account by making adjustments for factors that are not reflected in the econometric analysis. Then, the ranked companies are each allocated to an expenditure band: A, B, C, D or E. Some companies have a much better cost performance than suggested by the model – these are banded as 'A' companies. Other companies' cost performance is not as good as the models suggest it should be and their actual expenditure is well above that estimated by the model – they are banded as 'E' companies (OFWAT, 1998). Table 2.1 suggests that some companies may have lower than expected expenditure in both areas – the A/A companies – while others have higher than expected expenditure in both areas – the E/E companies. There are also companies with low expenditure in one area but high expenditure in the other.

These performance bands are used to set the 'catch-up' factors for each company. A frontier company is used to assess catching up, which is actually in line with the COLS method. The results are then used to set company-specific X factors in a price-cap formula. The X factor is higher for the relatively inefficient companies than for the efficient ones. Each company thus has its own price caps for five-year intervals. For the 2000-05 and 2005-10 periods the operating expenditure catch-up factor assumes that a water service company will catch up 60% of the assumed efficiency gap from its current performance to the frontier performance (i.e. a company-specific improvement). On top of that, the frontier shift of the continuing efficiency improvement factor (i.e. industry-wide or minimum efficiency improvement) is assumed. The capital-maintenance catch-up assumes reducing the efficiency gap by 40% in the case of cost-base comparisons and 50% in the case of estimates resulting from econometric analysis. In addition, the minimum efficiency improvement is set (OFWAT, 1999 and OFWAT, 2004).

Service performance adjustments are also taken into account when setting the price limits. Where the standard of service is assessed as being significantly better than that provided by the industry generally, an increase in price limits is made, whereas where the service is particularly poor relative to the industry a reduction is imposed (OFWAT, 1999).

Table 2.1: Matrix pattern for operating and capital-maintenance costs and the corresponding ranking of water supply companies

OPEX and CAPEX ranking		**Capital-maintenance expenditure**				
		A Less than 85% of C	**B** 85–95% of C	**C** Within 5% of modelled	**D** 105–115% of C	**E** More than 115% of C
Operating expenditure	**A** Less than 85% of C	Low	Low	Low	Low/high	Low/high
	B 85–95% of C	Low	Low	As expected	As expected	Low/high
	C Within 5% of modelled	Low	As expected	As expected	As expected	High
	D 105–115% of C	High/low	As expected	As expected	High	High
	E More than 115% of C	High/low	High/low	High	High	High

Source: OFWAT, 1998

All these estimates result in annual price limits.[18] It should be noticed that the total scope for efficiency improvement is not assumed in the price limit. In this way the companies can retain the benefits of outperforming for at least five years. After this five-year period prices are recalculated and the benefits of any improved performance are passed on to the final consumers.

2.3.3 Price Regulation in Italy

The Italian water industry is composed of approximately 6,000 companies and is highly fragmented. For instance, there are water companies serving less than 5,000 customers and companies serving more than 300,000 customers. Some water companies operate at the provincial level, whereas others operate at the municipal level. These companies are mostly public; there are only a few cases of private companies in this sector (Antonioli and Filippini, 2001). Sewage collection is nearly always operated by the municipality or joint boards of municipalities or other local authorities and is almost never integrated with water supply. In the past, in most cases tariffs used to be only sufficient to cover the operating costs of water supply and sewerage activities. Only in a few cases revenues resulting from tariffs allowed for new investments, and even for the depreciation of capital and maintenance. Investments were therefore typically financed from the public budget which provided funds in the form of subsidies or central grants (Massarutto, 1999).

In 1994, the Italian water sector underwent (regulatory) reforms with the aim to curtail local budget deficits. Law 36/1994, also known as the 'Galli law', aimed at a comprehensive reorganisation of water supply and sewerage services. The reform recognised the importance of economies of scale and introduced the full-cost pricing principle where the state sets the rules regarding the tariff structure for water prices and maximum increase rates set. In addition, with this regulatory reform the central government wanted to promote cost efficiency in the water sector by incorporating benchmarking in the price regulation process (Massarutto, 1999).

The tariffs of Italian water distribution companies are set at the local level. This autonomy is exercised within a general framework of rules set up at the national level concerning tariff structures, pricing criteria, obligations and above all maximum increase rates. In the field of water services these regulatory tasks are exercised by the CIPE (Interministerial Committee

[18] For example, the average price reduction of 2,8% per annum (before inflation) was set for the 2000-05 period (OFWAT, 1999).

for Economic Programming). The CIPE has the following responsibilities (Massarutto, 1999):

- obligation to achieve full-cost recovery;
- definition of cost that would include operating costs, maintenance costs and the costs of capital; and
- price increases should be determined according to the incentive-based regulatory scheme.

Up until 1995, maximum annual increase rates for water supply prices were set in line with the national policies concerning inflation. From 1996, maximum price increases have been set, allowing further increases for financing investments and in the case of municipalities at risk of bankruptcy. In 1999, the obligation to achieve at least 80% of cost recovery was added. This charging system was just an intermediate regime until full implementation of the 'Galli law' was achieved and a new price regulation system based on the rate-of-return regulation introduced, where eligible costs are specified in a detailed way. A rate of return on capital of 7% is foreseen. This new regime also introduces the benchmarking of operating costs based on a yardstick approach (Massarutto, 1999).

In practice, each firm defines its own tariff composed of a fixed charge and a variable component and submits this tariff to the regulation authority for approval. The tariff is approved only if the level of the variable component does not exceed a range of approximately 30% with respect to the benchmarking value obtained using the estimation results of a variable cost function. To correct the yardstick for the heterogeneity of the production process of water companies, the regulator uses the estimation results of a multivariate variable cost function. The parametric variable cost function for water distribution (also called '*Metodo Tariffario Normalizzato*,' MTN) that the Italian Regulation Authority proposes for calculating and evaluating the tariff is the following (Antonioli and Filippini, 2001):[19]

$$COAP = 0{,}9VE^{0{,}69} \times L^{0{,}33} \times IT^{0{,}1} \times e^{0{,}2\frac{Utdm}{UtT}} + EE + AA \qquad (2.7)$$

where:
COAP – the operating expenditure (million lire/year; after 2002 the currency is EUR)
VE – the volume of water delivered (thousand m^3/year)

[19] The parameters of this mathematical expression have been obtained by estimating a variable cost function for a sample of 20 companies for the year 1991.

L – length of the distribution network (km)
Utdm – the measured volume of water delivered to households
UtT – the total number of consumers (sum of household and non-household users)
EE – the electricity expenditure (million lire/year)
AA –expenditure on water bought (million lire/year)
IT – index for the difficulty of water treatment (in the absence of treatment, *IT* = 1)

Hence, the Italian Regulation Authority employs benchmarking to determine the cost range of the country's water distribution utilities. There is also a price limit so a price increase cannot exceed a certain limit calculated as a function of a starting level. The coefficient estimates in Eq.(2.7) may, however, be arguable since different estimates were obtained in a recalculation of the model using a panel data set by Antonioli and Filippini (2001).

2.4 Current Price Regulation Design of the Slovenian Water Sector

The Law on Price Control (1999) and more specifically the Decree on the Price Determination of Communal Services (2005) and Rules on the Price Determination of Obligatory Local Public Utilities for Environmental Protection (2004) form the existing regulatory framework for the price regulation of communal utilities in Slovenia, including water distribution utilities.

By the Decree on the Price Determination of Communal Services (2005), the price-setting mechanism is defined for obligatory local public utilities for environmental protection (also referred to as communal services). The price of the communal public service is to be determined *separately* for the following services: (i) supply of drinking water; (ii) drainage of wastewater; (iii) cleaning of wastewater; (iv) collection and transportation of municipal solid waste; (v) processing of solid waste; and (vi) the disposal of solid waste. The price of the communal public service (*P*), which does not include taxes or other levies, is defined by the following formula:

$$P = P_{TC} + P_I \,. \tag{2.8}$$

The price of the communal public service (*P*) is composed of two parts, the part to cover total operating and capital costs (P_{TC}) and the part that would be used to finance new infrastructure investments (P_I). Price is expressed in monetary units per physical unit of

service provided to customers (e.g., in SIT/m^3 for drinking water supply). If different users cause different costs to the provider of a communal public service, the price can be differentiated across different users or groups of users. However, this is not the case if the difference in costs is a result of different costs associated with access to the public service. The full-cost price is calculated as follows:

$$P_{TC} = \frac{Rev}{Q} = \frac{\sum_i C_i}{Q}, \tag{2.9}$$

where Rev is the required annual revenue, Q is the total physical quantity of service provided in one year (e.g., total amount of water supplied), and C_i is the i-th cost item, the sum of all cost items being equal to total annual cost (TC). The list of cost items C_i included in the calculation of full-cost price P_{TC} is provided in Table 2.2. The price is generally set in a way so as to cover the eligible operation and maintenance costs and depreciation.

P_I is that part of the price that should enable new investments in the infrastructure of a given public utility (i.e., building new infrastructure objects and facilities). The average price P_I is calculated as the investment cost per unit of service provided, divided by the number of years in which investment would be covered through the price of the public communal service. P_I excludes those costs for which funds are already provided from other sources (e.g., the national or local community budget, regional state aid, funds of the EU). P_I also excludes costs already taken into the account in the calculation of the full-cost price P_{TC} (e.g., profit intended for further development of the utility (see Table 2.2, category 5), unused net cash flow from previous years).

Under the Decree public utilities can increase part of their price intended to finance infrastructure investments, provided they have previously obtained the approval of Ministry of the Economy and provided that investments are planned in National Operative Programmes of Environmental Protection and local community development programmes up until the year 2008. The proposed increase must not exceed the increase in the industrial producer price index in the period from January 2004 until the date the application is filed. Further, an increase in P_I has to take into account potential cost savings and additional revenues (due to an increased number of customers) as a result of new investments. The price P_I can also be increased if the utility is faced with a difficult financial situation due to low prices not allowed to increase in previous years. The increase in the full-cost price P_{TC} is allowed only if there are objective and justified reasons for an increase in costs as a result of meeting the required standards, or if new services are introduced in accordance with

environmental regulations. Again, the approval of the Ministry of the Economy is needed. The application to be filed for obtaining such approval is very detailed in nature. If the final price for communal public service *P* were to increase by more than 2%, then the approval of the Ministry of Finance is also required.

Table 2.2: Specification of cost items included in the calculation of P_{TC} and the required revenue calculation

Nr.	Cost category
1	***Direct operating costs***
1.1	*Electrical energy*
1.2	*Fuel*
1.3	*Material*
1.4	*Services*
1.5	*Labour*
1.6	*Direct cost of sales*
1.7	*Other direct operating costs*
2	***Indirect operating costs***
2.1	*Depreciation*
2.2	*Maintenance*
2.3	*Other indirect operating costs*
3	***General costs***
3.1	*procurement*
3.2	*overhead and administrative costs*
3.3	*sales*
3.4	*Interest on debt capital*
4	***Total cost (TC)*** [= 1 + 2 + 3]
5	***Profit intended for utility's further development***
6	***Required revenue*** [= 4 + 5]

Source: Decree on the Price Determination of Communal Services (2005)

It can be concluded that at present the price regulation of Slovenian water distribution utilities largely resembles the rate-of-return regulation. This is combined with a very restrictive policy with respect to allowing price increases whereby the primary objective is to keep the inflation rate down rather than to influence the performance of water utilities. As long as the Decree on the Price Determination of Communal Services (2005) is valid, the Rules on Price Determination of Obligatory Local Public Utilities for Environmental Protection (2004) will not be applied. In fact, the Decree was introduced since it had been realised that introduction of the Rules was too ambitious an objective for the time being. Probably the key novelty of the Rules is the introduction of best practice and benchmarking

(in Article 8). According to the Rules, the full-cost price of the best practice utility P_{TC} is to be determined by the Ministry of the Environment and Spatial Planning based on results obtained via a questionnaire on costs and the quality of providing communal public services in representative areas. If the quality of communal public services does not reach the quality of the best practice provider, the full-cost price should be lower than in the best practice case. Comparative analysis or benchmarking should be performed to establish the cost of the best-practice utility. The comparative analysis is set to take into account the size of the utility (in terms of the quantity of services supplied) and thus control for any possible presence of economies of scale. For each group of comparable utilities in terms of size, the highest allowed deviation from average costs is also going to be defined. The Ministry can allow for a higher cost of providing communal public services if the public utility files an application containing an economic and technical analysis, from which it is evident that higher costs are the results of unfavourable natural conditions prevailing in the supply area, a low population density, a reduced efficiency of existing infrastructure facilities, or higher electricity and fuel costs. Until the best-practice operation is not defined, the full-cost price is going to be determined as the average cost of utilities providing communal public services.

The Rules on the Price Determination of Obligatory Local Public Utilities for Environmental Protection (2004) thus envisage the use of benchmarking methods in the present regulatory scheme based on rate-of-return regulation. In this way, utilities would be given an incentive for more efficient production as is already the case in the UK and Italian water industries. The Water Pricing Communication (COM(2000) 477 final) also facilitates the use of benchmarking. However, it is not yet certain when the Rules will come into effect. While the benchmarking method and best-practice performance with respect to the cost and required quality standards for carrying out public services have not yet been determined, implementation of the Rules was postponed by introducing the Decree on the Price Determination of Communal Services (2005). With respect to the choice of benchmarking methods, the regulatory authorities can decide to implement a simple comparison of one-dimensional measures of performance (i.e., performance indicators) or decide on more sophisticated benchmarking methods. In the thesis, we consider the possibility of employing parametric frontier benchmarking methods in the price regulation of Slovenian water distribution utilities. Our main objective is to obtain preliminary estimates of the cost inefficiency of Slovenian water distribution utilities and to establish whether the results obtained can be reliably used for price regulation. Therefore, several different parametric frontier methods will be employed in an estimation of the cost frontier function. We additionally consider alternative uses of the results since the estimated cost (frontier) function can also be used by regulatory authorities to predict utilities' costs and to estimate and decompose total factor productivity (TFP) growth.

3 Introduction to Efficiency Analysis

Classical economic theory predicts that firms or producers seek to maximise profit or minimise costs and to thus operate efficiently. However, evidence from practice does not always support this. Some firms tend to deviate from the predicted behaviour and are hence regarded as inefficient.[20] This required the development of a new line of theory that explicitly takes inefficiency into account. Modern efficiency measurement began with Farrell (1957) who drew upon the work of Debreu (1951) and Koopmans (1951) to define a simple measure of firm efficiency that could account for multiple inputs and be easily generalised to multiple outputs. In this chapter we introduce the theoretical tools needed to define different efficiency concepts. The tools and concepts used are derived directly from production theory. Thus, we start by briefly providing the analytical foundation of production theory. A more detailed discussion can be found in Chambers (1988), Cornes (1992), Mas-Colell et al. (1995) and Kumbhakar and Lovell (2000). The general case of a multi-output, multi-input production technology is considered. We first utilise information on the physical quantities of inputs and outputs to define distance functions, which provide the boundary of the production possibility set. As suggested by its name, the distance function gives measures of the distance of a firm's production activity to the boundary of the technology set, and is closely related to the concept of technical efficiency. Then, duality theory is employed to obtain an economic representation of the production possibility set. By considering input prices and under suitable behavioural assumptions the cost frontier can be defined. The cost frontier is then used as a standard (a benchmark, 'best practice' performance) against which cost efficiency can be measured. In particular, special stress is put on the cost frontier and the associated concept of cost efficiency since these two concepts are key to the cost efficiency analysis in the empirical part of the thesis. The optimal size of the firm and corresponding economies of scale and scale efficiency are also considered. Cost subadditivity is introduced as a necessary and sufficient condition for a natural monopoly. Further, a distinction between economies of scale and size is made. Besides economies of size, economies of output and customer density are considered as well since they play an important role in network industries.

[20] For example, public utility companies operating in network industries typically operate as national or local monopolies. Since they are not faced with competitive pressures they may not have sufficient incentives to operate efficiently.

3.1 Production Technology and Input Requirement Sets

The production possibility set (Chambers, 1988), production set (Mas-Colell et al., 1995), technology set (Coelli et al., 1998), or structure of production technology (Kumbhakar and Lovell, 2000) is initially described in terms of feasible sets of inputs and outputs. Let us assume that firms use a non-negative $K\times1$ vector of inputs, denoted by $\mathbf{x}$, to provide a non-negative $M\times1$ vector of outputs, denoted by $\mathbf{y}$. The technology set T represents the set of feasible input-output combinations given the existing state of technology:[21]

$$T = \{(\mathbf{x}, \mathbf{y}) : \mathbf{x} \text{ can produce } \mathbf{y}\}. \tag{3.1}$$

The technology set T satisfies the following properties:[22]

1. T is nonempty;
2. T is a closed set;
3. T is a convex set;[23]
4. T is bounded from above for every finite $\mathbf{x}$;
5. Weak disposability of $\mathbf{x}$: if $(\mathbf{x},\mathbf{y})\in T \;\Rightarrow\; (\lambda\mathbf{x},\mathbf{y})\in T \quad \text{for} \quad \lambda\geq 1$;
6. Weak disposability of $\mathbf{y}$: if $(\mathbf{x},\mathbf{y})\in T \;\Rightarrow\; (\mathbf{x},\lambda\mathbf{y})\in T \quad \text{for} \quad 0\leq\lambda\leq 1$;
7. $(0, \mathbf{x}) \in T$; and $(\mathbf{y}, 0) \in T \Rightarrow \mathbf{y} = 0$.

Property 1 says that a technology exists, i.e. outputs can be produced using inputs. Otherwise, there would be no need to study the behaviour of firms. Property 2 requires that the set T includes its boundary. It guarantees the existence of technically efficient input and output vectors. Property 3 implies that if $\mathbf{y}, \mathbf{y}' \in T$ and $\theta \in [0,1]$, then $\theta\mathbf{y} + (1-\theta)\mathbf{y}' \in T$. This convexity assumption has two important implications. Firstly, it implies nonincreasing returns to scale and, secondly, it captures the idea that 'unbalanced' input combinations are not more productive than 'balanced' ones. In particular, if production plans $\mathbf{y}$ and $\mathbf{y}'$ produce

[21] As an alternative to Eq.(3.1), T may be also defined in terms of $(K+M)$-dimensional vector $\mathbf{z}$ that contains both inputs and outputs. By convention, positive numbers denote outputs and negative numbers denote inputs. Production vector $\mathbf{z}$ is usually called input-output or netput vector, or a production plan (Mas-Colell et al., 1995). However, in what follows it is more convenient to maintain a clear dichotomy between inputs and outputs.

[22] The properties will not be proven here, we only explain their essence. Interested readers are referred to the literature mentioned earlier.

[23] T is not generally required to be a convex set. This property is required occasionally.

exactly the same amount of output but use different input combinations then a production vector that uses the average of the input vectors used in these two production plans can do at least as well as either $\mathbf{y}$ or $\mathbf{y}'$. Property 4 guarantees that finite input cannot produce infinite output. It is a mathematical regularity condition that guarantees the existence of a well-defined extreme for the optimisation problem. Properties 5 and 6 are weak monotonicity properties that guarantee the feasibility of *radial* expansions of feasible inputs and *radial* contractions of feasible outputs. These two properties can be replaced with the single strong monotonicity property given by: $(\mathbf{x},\mathbf{y}) \in T \;\Rightarrow\; (\mathbf{x}',\mathbf{y}') \in T,\ \forall (-\mathbf{x}',\mathbf{y}') \le (-\mathbf{x},\mathbf{y})$. If $\mathbf{x}$ can produce a given production bundle $\mathbf{y}$, then a bigger input bundle can also produce a given output bundle. Moreover, if $\mathbf{x}$ can produce $\mathbf{y}$ it can also produce all smaller output bundles. In this way, we do not limit ourselves to radial expansions or contractions only but guarantee the feasibility of *any* increase in feasible inputs and *any* reduction in feasible outputs. This property is also known as a strong or free disposability property. Finally, Property 7 says that any nonnegative input vector can produce at least a zero output, i.e. the origin belongs to T, and that it is not possible to produce something from nothing ('no free lunch' property).

An important characterisation of the technology set T is provided by input requirement sets, which is given in the following definition. The input requirement sets describe the sets $L(\mathbf{y})$ of input vectors $\mathbf{x}$ that are feasible for each output vector $\mathbf{y}$:[24]

$$L(\mathbf{y}) = \{\mathbf{x} : (\mathbf{x},\mathbf{y}) \in T\},\ \forall\ \mathbf{y}. \tag{3.2}$$

From the properties of technology set T it follows that input sets $L(\mathbf{y})$ satisfy the following properties:

1. $L(\mathbf{y})$ is nonempty for at least one finite output vector;[25]
2. the sets $L(\mathbf{y})$ are closed;
3. $L(\mathbf{y})$ is convex;
4. $\mathbf{x}$ is finite $\Rightarrow\ \mathbf{x} \notin L(\mathbf{y})$ if $\mathbf{y}$ is infinite;
5. If $\mathbf{x} \in L(\mathbf{y}) \;\Rightarrow\; \lambda\mathbf{x} \in L(\mathbf{y})$ for $\lambda \ge 1$;
6. $L(\mathbf{y}) \subseteq L(\lambda\mathbf{y})$ for $0 \le \lambda \le 1$;

[24] The input requirement set (Chambers, 1988) is also referred to as the input set (Coelli et al., 1998) or the input set of production technology (Kumbhakar and Lovell, 2000).

[25] Property 1 of T implies that at least one feasible input-output combination exists. It does not, however, imply that $L(\mathbf{y})$ is nonempty for all $\mathbf{y}$.

7. $0 \notin L(\mathbf{y})$ for $\mathbf{y} \geq 0$; and $L(0) = \Re_+^K$.

Again, if the weak monotonicity Properties 5 and 6 are replaced with the strong monotonicity property we get: $\mathbf{x} \in L(\mathbf{y}), \ \mathbf{x}' \geq \mathbf{x} \ \Rightarrow \ \mathbf{x}' \in L(\mathbf{y})$ and $\mathbf{y}' \leq \mathbf{y} \ \Rightarrow \ L(\mathbf{y}) \subseteq L(\mathbf{y}')$. The second part of this property says that if $\mathbf{x}$ can produce $\mathbf{y}$ it can also produce a smaller output bundle $\mathbf{y}'$, while knowing that $\mathbf{x}'$ can produce $\mathbf{y}'$ does not necessary mean that it can also produce a bigger output bundle $\mathbf{y}$. Property 7 indicates that non-zero output levels cannot be produced from zero input levels and that inaction is possible, i.e. nothing can be produced out of a given set of inputs.

We now focus on the boundaries of the input requirement sets $L(\mathbf{y})$, namely input isoquants and input efficient subsets, which play an important role in the efficiency analysis. The input isoquants are defined in the following way:

$$Isoq\, L(\mathbf{y}) = \{\mathbf{x} : \mathbf{x} \in L(\mathbf{y}), \lambda \mathbf{x} \notin L(\mathbf{y}), \lambda < 1\}. \tag{3.3}$$

Further, the input efficient subsets are defined as:

$$Eff\ L(\mathbf{y}) = \{\mathbf{x} : \mathbf{x} \in L(\mathbf{y}), \mathbf{x}' \leq \mathbf{x} \ \Rightarrow \ \mathbf{x}' \notin L(\mathbf{y})\}. \tag{3.4}$$

The input isoquants describe the sets of input vectors capable of producing each output vector $\mathbf{y}$ but which, when *radially* contracted, become incapable of producing given output vector $\mathbf{y}$. Alternatively, the input efficient subsets describe the sets of input vectors capable of producing each output vector $\mathbf{y}$ but which, when contracted in *any* dimension, become incapable of producing given output vector $\mathbf{y}$. *Isoq* $L(\mathbf{y})$ represents one notion of minimal input use and appears to provide an appealing standard against which to measure technical efficiency. Yet, *Isoq* $L(\mathbf{y})$ can also include input vectors $\mathbf{x}$ that belong to the uneconomic region of input space.[26] On the other hand, *Eff* $L(\mathbf{y})$ only includes input vectors belonging to the economic region of input space and it holds that *Eff* $L(\mathbf{y}) \subseteq$ *Isoq* $L(\mathbf{y})$.[27] Thus, *Eff* $L(\mathbf{y})$ provides a more stringent standard against which to measure the technical efficiency of input use.

[26] In the two-input space this can be represented by an upward sloping isoquant.

[27] Some functional forms employed in econometric analysis such as Cobb-Douglas do have the property that *Eff* $L(\mathbf{y})$ = *Isoq* $L(\mathbf{y})$, making the distinction irrelevant. Other functions, such as translog, have the property that *Eff* $L(\mathbf{y}) \subset$ *Isoq* $L(\mathbf{y})$, making the distinction potentially important (Kumbhakar and Lovell, 2000).

3.2 Distance Functions

We now turn to distance functions which provide a functional characterisation of the production technology and were introduced into economic theory by Shephard (1953, 1970). More precisely, distance functions allow the specification of a multiple-input, multiple-output production technology. Alternatively, a multiple-input multiple-output production technology may be specified by the production possibility function or production transformation function, $F(\mathbf{y}, \mathbf{x}) = 0$, which provides the boundary of production possibility set when multiple inputs are used to produce multiple outputs. In what follows, this concept will not be examined any further since it is rarely used in the empirical research. Moreover, distance functions prove to be far more useful and straightforward in defining technical efficiency, as will be seen later in this chapter.[28] Further, our main focus in the empirical analysis will be on the cost function and cost efficiency rather than on the production function and technical efficiency. For the purpose of our analysis it thus suffices to just introduce the concept of distance functions.

A distance function may have either an input or an output orientation. An input distance function characterises the production technology by looking at the maximal proportional contraction of the input vector, given an output vector. An output orientation looks at how much the output vector may be proportionally expanded with the input vector being held fixed. Given that most public utility companies have an obligation to meet demand, they can only become more efficient by providing a predefined output level with fewer inputs. We therefore utilise an input orientation approach in what follows.

The biggest role distance functions play is in duality theory. It can be shown that in certain conditions an input distance function is dual to a cost (frontier) function. Nevertheless, distance functions are not without their empirical value. They can be utilised to obtain measures of technical efficiency when firms use multiple inputs to produce multiple outputs. The input distance function is a function:

$$D_I(\mathbf{x}, \mathbf{y}) = \max\{\rho : (\mathbf{x}/\rho) \in L(\mathbf{y})\} \ . \tag{3.5}$$

[28] In a single output case, technical efficiency can be as well easily defined using the production function: $f(\mathbf{x}) = \max\{y : \mathbf{x} \in L(y)\}$. However, we will leave the discussion of the production function aside since we do not wish to limit our attention to a single output case but want to keep the discussion as general as possible.

The input distance function gives the maximum amount by which a producer's input vector can be *radially* contracted and still remain feasible for the output vector it produces. Since the input distance function $D_I(\mathbf{x}, \mathbf{y})$ is defined in terms of the input sets $L(\mathbf{y})$, the properties of $D_I(\mathbf{x}, \mathbf{y})$ can be easily derived from the properties of $L(\mathbf{y})$. The input distance function thus satisfies a corresponding set of properties:

1. $\mathbf{x} \in L(\mathbf{y}) \quad \Rightarrow \quad D_I(\mathbf{x}, \mathbf{y}) \geq 1$;
2. $D_I(\mathbf{x}, \mathbf{y})$ is homogeneous of degree one (or, linearly homogeneous) in $\mathbf{x}$: $D_I(\lambda \mathbf{x}, y) = \lambda D_I(\mathbf{x}, \mathbf{y})$, for $\lambda > 0$;
3. $D_I(\mathbf{x}, \mathbf{y})$ is a concave function in $\mathbf{x}$;
4. $D_I(\mathbf{x}, \mathbf{y})$ is an upper-semicontinuous function;
5. $D_I(\lambda \mathbf{x}, \mathbf{y}) \geq D_I(\mathbf{x}, \mathbf{y})$ for $\lambda \geq 1$;
6. $D_I(\mathbf{x}, \lambda \mathbf{y}) \leq D_I(\mathbf{x}, \mathbf{y})$ for $\lambda \geq 1$.

When a firm is technically efficient, $D_I(\mathbf{x}, \mathbf{y}) = 1$. Clearly, we can rewrite the definition for input requirement sets as $L(\mathbf{y}) = \{\mathbf{x} : D_I(\mathbf{x}, \mathbf{y}) \geq 1\}$, and respectively the definition for input isoquants as $Isoq\, L(\mathbf{y}) = \{\mathbf{x} : D_I(\mathbf{x}, \mathbf{y}) = 1\}$. The input isoquant corresponds to the set of input vectors having the value of input distance function equal to unity and thus being technically efficient. All other feasible input vectors have input distance function values greater than unity.

3.3 Cost Functions

So far, we have only used information on the quantities of inputs and outputs to describe the production technology. To obtain a cost function, we must introduce input prices and specify a behavioural objective. We thus assume that companies face a strictly positive $K \times 1$ vector of input prices $\mathbf{w}$ and that they tend to minimise the cost of producing the chosen output vector $\mathbf{y}$. Another assumption regarding prices is that the producers are price-takers in input markets meaning that they do not have sufficient market power to influence the prices and thus take the prices as given. The existing literature on cost-minimising behaviour and cost functions is fairly extensive. The theory presented in the following section is based on Varian (1984), Chambers (1988), Jehle and Reny (1998) and Mas-Colell et al. (1995). The proofs will not be presented here. Interested readers are referred to the abovementioned literature.

Either the input sets or the input distance function can be used to derive a cost function. The cost function is defined as follows:

$$\begin{aligned} C(\mathbf{y},\mathbf{w}) &= \min_{\mathbf{x}}\left\{\mathbf{w}^{\mathrm{T}}\mathbf{x} : \mathbf{x} \in L(\mathbf{y})\right\} \\ &= \min_{\mathbf{x}}\left\{\mathbf{w}^{\mathrm{T}}\mathbf{x} : D_I(\mathbf{x},\mathbf{y}) \geq 1\right\}. \end{aligned} \tag{3.6}$$

The cost function achieves the minimum expenditure required to produce any output vector, given input prices and production technology. It should be noted that the cost function defined above is in line with the neoclassical microeconomic theory where it is assumed that all firms minimise costs and are therefore cost efficient. Nevertheless, according to the evidence from practice some firms may fail to minimise their costs and are found to be cost inefficient. For example, if firms operate in non-competitive environment that does not provide sufficient incentives for efficient production managerial behaviour may not be consistent with the cost-minimising pattern. In such cases, inefficiency has to be explicitly taken in the account when modelling firms' behaviour. Therefore, an empirical concept of the cost frontier function was developed as opposed to a purely theoretical concept of the cost function.[29] Introduction of the cost frontier function can be viewed as a reflection of the fact that in practice some firms fail to attain the frontier and therefore we have to admit some inefficiency, i.e. the possibility of systematic divergence between observed and minimal costs. We turn to this and some related issues in Section 3.4 where different efficiency concepts are discussed.

Properties of the cost function can be derived from the properties of the input sets and the input distance function. Hence, the cost function satisfies the following properties:

1. $C(\mathbf{y},\mathbf{w}) > 0$ for $\mathbf{y} \geq 0$ (nonnegativity) and $C(0,\mathbf{w}) = 0$ (no fixed costs);
2. $C(\mathbf{y},\lambda\mathbf{w}) = \lambda C(\mathbf{y},\mathbf{w})$ for $\lambda > 0$ (linear homogeneity);
3. $C(\mathbf{y},\mathbf{w}') \geq C(\mathbf{y},\mathbf{w})$ for $\mathbf{w}' \geq \mathbf{w}$ (nondecreasing in **w**);
4. C(**y**, **w**) is a concave and continuous function in **w**;
5. $C(\lambda\mathbf{y},\mathbf{w}) \geq C(\mathbf{y},\mathbf{w})$ for $\lambda \geq 1$ (nondecreasing in **y**);[30]

[29] Another empirical concept is the average cost function, where we do not allow for inefficiency. All deviations from the estimated cost function are attributed to random noise.

[30] If, in addition, we would like C(**y**, **w**) to be differentiable in **y** for **y** > 0, we have to assume that C(**y**, **w**) is lower semicontinuous in **y**. This simply says that marginal costs exist for positive levels of

6. T is a convex set $\Rightarrow C(\mathbf{y}, \mathbf{w})$ is a convex function in $\mathbf{y}$.

Property 1 states that it is impossible to produce a positive output with no costs. This is a consequence of the fact that prices are assumed to be strictly positive and at least one input is required to produce an output. The no fixed costs property implies that it is costless to produce a zero output.[31] Property 2 is called the linear homogeneity property and says that, when all prices change proportionally, then total costs will also change in the same manner (they will be increased or reduced by the same proportionality factor λ). This is restatement of the principle that only relative prices matter to economic agents. As long as the input prices vary only proportionately, the cost-minimising choice of inputs will not vary. Property 3 indicates that costs will increase when at least one input price rises and the others stay the same. According to Property 4, when input prices increase the costs increase at most in a linear way. This is because of the substitution effect which allows firms to change the relative use of different inputs if input prices change. If substitution is technologically not possible, the costs rise linearly.[32] Continuity is needed in order to be able to calculate partial derivatives with respect to w_k ($k = 1, \ldots, K$). Property 5 is a weak monotonicity property and can be replaced by the strong monotonicity property expressed by: $C(\mathbf{y}', \mathbf{w}) \geq C(\mathbf{y}, \mathbf{w})$ for $\mathbf{y}' \geq \mathbf{y}$. It says that costs cannot decrease as output increases. If T is convex, then Property 6 holds also so that $C(\mathbf{y}, \mathbf{w})$ is a convex function in $\mathbf{y}$.

Under Properties 1–4, the strong monotonicity property, and Property 6, the cost function $C(\mathbf{y}, \mathbf{w})$ is dual to the input distance function $D_I(\mathbf{y}, \mathbf{x})$, i.e. $C(\mathbf{y}, \mathbf{w})$ and $D_I(\mathbf{y}, \mathbf{x})$ provide equivalent representations of the technology set on the assumption of cost-minimising behaviour and in the presence of exogenously determined input prices. As will be seen in Section 3.4, the duality relationship linking a cost frontier with an input distance function is proven to be important for the measurement and decomposition of cost efficiency.

If the cost function $C(\mathbf{y}, \mathbf{w})$ is in addition differentiable with respect to input prices $\mathbf{w}$, then there exists a unique vector of cost-minimising input demand equations that is equal to the gradient of $C(\mathbf{y}, \mathbf{w})$ in input prices $\mathbf{w}$. That is, if $\mathbf{x}(\mathbf{y}, \mathbf{w})$ is a unique vector of cost-minimising input demand equations then Shephard's (1953) lemma states that:

---

output. We allow for the possibility of start-up resource needs that would give $C(\mathbf{y}, \mathbf{w})$ a jump discontinuity at $\mathbf{y} = 0$.

[31] This holds when we are dealing with a long-run problem where all inputs are perfectly variable. A discussion of long-run vs. short-run costs follows in Section 3.3.2.

[32] The Leontief production function is such an example.

$$\mathbf{x}(\mathbf{y},\mathbf{w}) = \nabla_{\mathbf{w}} C(\mathbf{y},\mathbf{w}) .^{33} \tag{3.7}$$

For a cost-minimising producer, $\mathbf{w}^{\mathrm{T}}\mathbf{x} = C(\mathbf{y}, \mathbf{w})$ and $\mathbf{x} = \mathbf{x}(\mathbf{y}, \mathbf{w})$. An important implication of Shephard's lemma is that the behaviour of derived input demands is determined by the cost function. Properties of the cost function place implicit conditions on the cost-minimising input demands. We proceed by studying these implications in more detail.

3.3.1 Input Demand Elasticity and Elasticity of Substitution

According to Shephard's (1953) lemma, the conditional input demand vector is equal to the gradient of the cost function in input prices. Using this lemma, we can express the cost shares of all inputs as elasticities of the cost function with respect to the input prices:

$$S_k(\mathbf{y},\mathbf{x}) = \frac{w_k x_k(\mathbf{y},\mathbf{w})}{C(\mathbf{y},\mathbf{w})} = \frac{\partial \ln C(\mathbf{y},\mathbf{w})}{\partial \ln w_k} . \tag{3.8}$$

It has been already established that when all input prices are increased proportionally costs increase in the same proportion. The linear homogeneity of the cost function then implies that the conditional factor demand functions are homogeneous of degree zero in input prices.[34] If input prices change proportionately, the conditional factor demands will not change, i.e. only relative prices matter. Mathematically, this is expressed as:

$$\mathbf{x}(\mathbf{y},t\mathbf{w}) = \mathbf{x}(\mathbf{y},\mathbf{w}) . \tag{3.9}$$

Further, it has already been implied that a rise in any input price causes a decline in use of that input. Consequently, conditional input demand curves must be downward sloping. Applying Shephard's lemma gives:

$$\frac{\partial x_k(\mathbf{y},\mathbf{w})}{\partial w_j} = \frac{\partial^2 C(\mathbf{y},\mathbf{w})}{\partial w_k \partial w_j}, \qquad k, j = 1, \ldots, K. \tag{3.10}$$

---

[33] A sufficient condition for the existence of a unique cost-minimising solution is that input requirement set $L(\mathbf{y})$ is a strictly convex set.

[34] Since the partial derivatives of a function homogeneous of degree k are homogeneous of degree k–1.

All conditional input demand responses to input prices can thus be computed directly from the Hessian matrix of the cost function. Concavity in input prices and twice-continuous differentiability of $C(\mathbf{y}, \mathbf{w})$ imply that the Hessian matrix $H_{\mathbf{ww}}\, C(\mathbf{y}, \mathbf{w})$ is negative semidefinite. Thus, $\partial x_k(\mathbf{y},\mathbf{w})/\partial w_k \le 0$, implying that as some input becomes more expensive you buy less of that input. The symmetry condition expressed as $\partial x_k(\mathbf{y},\mathbf{w})/\partial w_j = \partial x_j(\mathbf{y},\mathbf{w})/\partial w_k$ is merely a mechanical implication of the presumed differentiability properties of the cost function; there is no economic intuition behind it.

Input responsiveness to changes in input prices can also be expressed by conditional input demand elasticity as follows:

$$\epsilon_{kj} = \frac{w_j}{x_k(\mathbf{y},\mathbf{w})}\frac{\partial x_k(\mathbf{y},\mathbf{w})}{\partial w_j}. \tag{3.11}$$

The homogeneity of degree zero according to Euler's theorem implies $\sum_j \left(\partial x_k(\mathbf{y},\mathbf{w})/\partial w_j\right) w_j = 0$, which combined with Eq.(3.11) yields $\Sigma_j\ \epsilon_{kj} = 0$, while the negative semidefiniteness of Eq.(3.10) and Eq.(3.11) implies $\epsilon_{kk} \le 0$. In general, these elasticities are not symmetric, i.e. $\epsilon_{kj} \ne \epsilon_{jk}$. However, it can be easily shown that:

$$\epsilon_{kj} = \frac{S_j}{S_k}\,\epsilon_{jk}, \tag{3.12}$$

where S_k and S_j are the cost shares of k-th and j-th input, respectively. Further, one can also be interested in relative input responsiveness to changes in relative input prices. This is measured by the elasticity of substitution σ which, in the two input case, can be written as follows:

$$\begin{aligned}\sigma &= \frac{d\ln(x_2/x_1)}{d\ln(w_1/w_2)} \\ &= \frac{d(x_2/x_1)}{(x_2/x_1)} \Big/ \frac{d(w_1/w_2)}{(w_1/w_2)}.\end{aligned} \tag{3.13}$$

The elasticity of substitution can be thus interpreted as the elasticity of the input ratio with respect to the input price ratio.[35] For convex isoquants it lies between 0 and ∞, with a larger value of σ implying greater substitutability between the inputs. A value of $\sigma = \infty$ occurs when the inputs are perfectly substitutable, while $\sigma = 0$ implies that no substitution is possible. With more than two inputs, however, Eq.(3.13) becomes more complex. A common definition is the Allen-Uzawa concept of the elasticity of substitution introduced by Allen (1938) and Uzawa (1962). This partial elasticity of substitution defines the elasticity of substitution for each pair of inputs as:

$$\sigma_{kj} = \frac{C(\mathbf{y},\mathbf{w})\, C_{kj}(\mathbf{y},\mathbf{w})}{C_k(\mathbf{y},\mathbf{w})\, C_j(\mathbf{y},\mathbf{w})}, \tag{3.14}$$

where subscripts related to the cost function refer to the partial derivative(s) with respect to the associated input price(s). It turns out that Eq.(3.14) can be expressed in terms of factor elasticities, ϵ_{kj}, and input factor shares, S_j:

$$\sigma_{kj} = \epsilon_{kj} / S_j \,. \tag{3.15}$$

In their paper Blackorby and Russell (1989) showed that the Allen elasticity of substitution (AES) is only an appropriate measure of substitution in specific cases and provides no additional information besides the factor elasticities and the factor shares. An alternative measure is the Morishima (1967) elasticity of substitution, which is defined as follows:

$$M_{kj} = \epsilon_{jk} - \epsilon_{kk}. \tag{3.16}$$

Blackorby and Russell (1989) showed that the Morishima elasticity of substitution (MES) preserves the important characteristics of the two-input elasticity in Eq.(3.13) and has several advantages over the (AES). They demonstrated that MES measures the curvature of an isoquant, it is a sufficient statistic for evaluating changes in relative prices and quantities, and it is a log derivative of the input quantity ratio with respect to the input price ratio. These

[35] The original definition states that elasticity of substitution equals the elasticity of input ratio with respect to the marginal product ratio, the latter being equal to marginal rate of technical substitution. Nevertheless, the first-order conditions for cost minimisation imply that the marginal rate of technical substitution between the *k*-th and *j*-th input equals the ratio of the *k*-th to the *j*-th input prices. A generalisation of the expression for the elasticity of substitution in more than two input case is provided in Chambers (1988).

characteristics do not apply to the AES. The MES is thus a more natural extension of the multi-input case. An important characteristic of the MES is its inherent asymmetry. Asymmetry appears to be natural as the partial derivative has to be evaluated in the direction of the input price that actually changes. For any cost function with more than two inputs, the MES is only symmetric in the special case where the cost function is of the constant elasticity of substitution (CES) type. The AES, on the other hand, is by definition symmetric for all input pairs which can be seen if we put Eq.(3.12) in Eq.(3.15).

3.3.2 Short vs. Long-Run Cost Function

Economists refer to short-run decisions as those that involve some fixity of inputs and long-run decisions as those that involve no fixed inputs. So far, all elements of input vector **x** have been treated as freely variable. Hence, the analysis has only encompassed long-run optimisation problems, with $C(\mathbf{y}, \mathbf{w})$ being a long-run (total) cost function. However, producers are sometimes faced with inputs that are only available in limited amounts. This creates an additional constraint in their cost-minimisation decision so the solutions to the short-run and long-run problems are not necessarily the same. Labour is typically considered to be a flexible input in the production process since a firm can easily alter the number of employees to arrive at its optimal level.[36] On the other hand, the capital of a firm is considered to be a fixed input since time is required to adjust it to its optimal level through the investment process. In what follows, short-run variable and total costs are presented and their relation to the long-run total costs is briefly discussed.

Suppose that the input vector is partitioned into two components with $\mathbf{x}_1$ containing perfectly variable inputs and $\mathbf{x}_2$ containing those inputs that are fixed or subject to some availability constraint. The restricted input requirement set is defined as:

$$L(\mathbf{y}, \mathbf{x}_2) = \{\mathbf{x}_1 : (\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) \in T\}. \tag{3.17}$$

The short-run variable cost function $VC(\mathbf{y}, \mathbf{w}_1, \mathbf{x}_2)$ is then defined as follows:

[36] In the real world, however, labour unions, job contracts (e.g. permanent employment) and labour legislation sometimes make it difficult to fire employees and the flexibility of labour can be questioned (i.e. employees may be considered a quasi-fixed input).

$$\begin{aligned} VC(\mathbf{y},\mathbf{w}_1,\mathbf{x}_2) &= \min_{\mathbf{x}_1}\left\{\mathbf{w}_1^{\mathrm{T}}\mathbf{x}_1 : \mathbf{x}_1 \in L(\mathbf{y},\mathbf{x}_2)\right\} \\ &= \min_{\mathbf{x}_1}\left\{\mathbf{w}_1^{\mathrm{T}}\mathbf{x}_1 : D_I(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}) \geq 1\right\}, \end{aligned} \tag{3.18}$$

where $\mathbf{w}_1$ is the set of variable input prices. A well-defined short-run variable-cost function satisfies the same properties as the long-run total-cost function in terms of $\mathbf{w}_1$ and $\mathbf{y}$. In addition to these properties, $VC(\mathbf{y}, \mathbf{w}_1, \mathbf{x}_2)$ satisfies another property that it is nonincreasing in $\mathbf{x}_2$:

$$\mathbf{x}_2 \geq \mathbf{x}_2^* \quad \Rightarrow \quad VC(\mathbf{y},\mathbf{w}_1,\mathbf{x}_2) \leq VC(\mathbf{y},\mathbf{w}_1,\mathbf{x}_2^*). \tag{3.19}$$

If the availability of $\mathbf{x}_2$ increases, new choices for $\mathbf{x}_1$ become feasible. This opens up new cost-minimising opportunities. Hence, variable costs cannot increase since what is the lowest cost now may not have been even available before the constraint was relaxed.

Since input $\mathbf{x}_2$ is presumably not free, no-fixed-cost property does obviously not hold in the short run. The short-run total cost function associated with producing the output vector $\mathbf{y}$ is:

$$C^S(\mathbf{y},\mathbf{w},\mathbf{x}_2) = VC(\mathbf{y},\mathbf{w}_1,\mathbf{x}_2) + \mathbf{w}_2^{\mathrm{T}}\mathbf{x}_2. \tag{3.20}$$

The main difference between $C(\mathbf{y}, \mathbf{w})$ and $C^S(\mathbf{y}, \mathbf{w}_1, \mathbf{x}_2)$ is that in Eq.(3.20) the fixed inputs do not necessarily minimise costs. However, by definition, variable costs are minimised for any given $\mathbf{x}_2$. The relationship between the short-run and long-run total cost function is then:

$$C(\mathbf{y},\mathbf{w}) = \min_{\mathbf{x}_2} VC(\mathbf{y},\mathbf{w}_1,\mathbf{x}_2) + \mathbf{w}_2^{\mathrm{T}}\mathbf{x}_2. \tag{3.21}$$

The long-run problem is decomposed into two components, that of minimising $VC(\mathbf{y}, \mathbf{w}_1, \mathbf{x}_1)$ given $\mathbf{x}_2$ and then choosing $\mathbf{x}_2$. Long-run total costs are thus equal to short-run total costs evaluated at the fixed input vector, which minimises long-run costs.

From Eq.(3.21) it follows that $C(\mathbf{y},\mathbf{w}) \leq C^S(\mathbf{y},\mathbf{w},\mathbf{x}_2)$ and $C(\mathbf{y},\mathbf{w}) = C^S\left(\mathbf{y},\mathbf{w},\mathbf{x}_2(\mathbf{y},\mathbf{w})\right)$, where $\mathbf{x}_2$ $(\mathbf{y}, \mathbf{w})$ is the solution of Eq.(3.21). These expressions imply that $C(\mathbf{y}, \mathbf{w})$ is the lower envelope of the respective $C^S(\mathbf{y}, \mathbf{w}_1, \mathbf{x}_1)$ functions in both input price and output space. It also follows that $C(\mathbf{y}, \mathbf{w})$ is more concave in $\mathbf{w}$ than $C^S(\mathbf{y}, \mathbf{w}_1, \mathbf{x}_1)$, meaning that long-run conditional factor demand is more own-price elastic than short-run conditional factor demand.

From the first-order condition associated with Eq.(3.21), the vector of shadow prices for the fixed input can be obtained as:

$$\mathbf{w}_2 = -\nabla_{\mathbf{x}_2} VC(\mathbf{y}, \mathbf{w}_1, \mathbf{x}_2). \tag{3.22}$$

Fixed inputs are acquired up to the point where the associated decrease in variable costs is just balanced by the marginal cost increment ($\mathbf{w}_2$).

The decision to utilise either a short-run or long-run cost function in the analysis is generally based on our belief whether the firms use all inputs at their optimal levels. If this is the case, then the total cost function is used in the analysis. On the contrary, if firms do not operate at their static equilibrium levels the variable cost function has to be employed. Moreover, in the empirical analysis the decision on variable as opposed to total cost is influenced by data availability and econometric considerations. In empirical studies it is often the case that a positive relationship between variable cost and the capital stock is found. There are two possible explanations of this theoretically implausible sign. The first interpretation suggested by Cowing and Holtmann (1983) argues that the positive sign of the coefficient of capital stock is an indicator of an excessive amount of capital stock being employed by firms. In this case, an increase in the capital stock would lead to an increase in both variable and total costs. According to the second interpretation, the incorrect sign of the coefficient of the capital stock is derived from the multicollinearity between the output and the capital stock (Guyomard and Vermersch, 1989, and Fillipini, 1996). This problem is often combined with the empirical difficulty of defining and measuring the capital stock variable. Due to a lack of data, most studies have used physical measures of the capital stock as proxies. These proxy variables are usually highly correlated with the output variable and may thus cause the multicollinearity problem (Fillipini, 1996).

3.4 Different Efficiency Concepts

In this section, an input distance function and a cost frontier function are utilised to introduce different efficiency measures. The efficiency measures are essentially defined in such a way as to provide measures of distance to a respective 'frontier' function (e.g. an input distance function or a cost frontier function). The definition of the cost frontier function corresponds to the definition of the cost function in Eq.(3.6). The only difference is that from now on observed costs are allowed to deviate from the minimum or frontier costs due to the cost inefficiency.

3.4.1 Technical Efficiency

According to Koopmans (1951), a producer is technically efficient if, and only if, it is impossible to produce more of any output without producing less of some other output or using more of some input. According to a formal definition of technical efficiency, an input-output vector $(\mathbf{x}, \mathbf{y}) \in T$ is technically efficient if, and only if, $(\mathbf{x}', \mathbf{y}') \notin T$ for $(-\mathbf{x}', \mathbf{y}') \geq (-\mathbf{x}, \mathbf{y})$.

An input-oriented definition of technical efficiency states that input vector $\mathbf{x} \in L(\mathbf{y})$ is technically efficient if, and only if, $\mathbf{x}' \notin L(\mathbf{y})$ for $\mathbf{x}' \leq \mathbf{x}$ or, equivalently, $\mathbf{x} \in \mathit{Eff}\, L(\mathbf{y})$. A feasible input vector is thus technically efficient if, and only if, no reduction in any input is feasible, holding the output vector fixed. Following this definition, we can define an input-oriented measure of technical efficiency first proposed by Debreu (1951) and Farrell (1957) as a function:

$$TE_I(\mathbf{x}, \mathbf{y}) = \min\{\theta : \theta \mathbf{x} \in L(\mathbf{y})\}. \tag{3.23}$$

Technical efficiency is measured in terms of an equi-proportional contraction of all inputs. If no such contraction is feasible, i.e. $TE_I(\mathbf{x}, \mathbf{y}) = 1$, then the input vector is technically efficient. It should be noted that equi-proportional contractions of inputs associate technical efficiency with membership in input isoquants, which is a necessary but not a sufficient condition for membership in input efficient subsets. Consequently, the above defined measure of technical efficiency is necessary but not sufficient for being technically efficient consistent with Koopmans' (1951) definition. However, since radial measures have nice technical properties a vast amount of the economic and econometric literature uses the Debreu (1951) and Farrell (1957) definitions. The input-oriented measure of technical efficiency given in Eq.(3.23) satisfies the following properties (Kumbhakar and Lovell, 2000):

1. $TE_I(\mathbf{x}, \mathbf{y}) \leq 1$;
2. $TE_I(\mathbf{x}, \mathbf{y}) = 1 \Leftrightarrow \mathbf{x} \in \mathit{Isoq}\, L(\mathbf{y})$;
3. $TE_I(\mathbf{x}, \mathbf{y})$ is nonincreasing in $\mathbf{x}$;
4. $TE_I(\mathbf{x}, \mathbf{y})$ is homogeneous of degree –1 in $\mathbf{x}$;
5. $TE_I(\mathbf{x}, \mathbf{y})$ is invariant with respect to the units in which $\mathbf{y}$ and $\mathbf{x}$ are measured.

Property 1 is a normalisation property which states that $TE_I(\mathbf{x}, \mathbf{y})$ is bounded above by unity. Property 2 states that $TE_I(\mathbf{x}, \mathbf{y})$ uses the relaxed standard $\mathit{Isoq}\, L(\mathbf{y})$ rather than the more

stringent standard *Eff L*(**y**) to measure technical efficiency. This is the only undesirable property of TE_I (**x**, **y**). Alternatively, technical efficiency can be defined relative to input efficient subsets but this would require replacing radial efficiency measures with nonradial efficiency measures where the latter do not satisfy Properties 4 and 5. Property 3 is a weak monotonicity property saying that TE_I (**x**, **y**) does not increase when the usage of any input increases. Property 4 is a homogeneity property saying that an equiproportionate increase in all inputs results in an equivalent change in an opposite direction in TE_I (**x**, **y**). Property 5 is an invariance property saying that efficiency scores are unaffected by changing the units in which any input or output is measured.

The input distance function is closely related to the input-oriented measure of technical efficiency which can be also written as: $TE_I(\mathbf{x},\mathbf{y}) = \min\{\theta : D_I(\theta\mathbf{x},\mathbf{y}) \geq 1\}$. Technical efficiency as defined above is equal to the inverse of the input distance function.

3.4.2 Cost Efficiency

While the standard against which technical efficiency is measured is provided by input isoquants, the cost frontier C(**y**, **w**) is an appropriate standard against which to measure cost (or overall) efficiency. In the first case, no behavioural objective needs to be specified whereas in the second case it is assumed that cost minimisation is an appropriate behavioural objective.

A measure of cost efficiency is the following function:

$$CE(\mathbf{x},\mathbf{y},\mathbf{w}) = C(\mathbf{y},\mathbf{w}) / \mathbf{w}^{\mathrm{T}}\mathbf{x}\,. \tag{3.24}$$

The cost frontier function is thus closely related to a measure of cost efficiency which is given by the ratio of minimum/frontier to observed cost. The properties satisfied by the measure of cost efficiency are (Kumbhakar and Lovell, 2000):

1. $0 < CE(\mathbf{x}, \mathbf{y}, \mathbf{w}) \leq 1$;
2. $CE(\mathbf{x}, \mathbf{y}, \mathbf{w}) = 1 \Leftrightarrow \mathbf{x} = \mathbf{x}(\mathbf{y}, \mathbf{w})$ and $\mathbf{w}^{\mathrm{T}}\mathbf{x} = C(\mathbf{y}, \mathbf{w})$;
3. $CE(\lambda\mathbf{x}, \mathbf{y}, \mathbf{w}) = \lambda^{-1} CE(\mathbf{x}, \mathbf{y}, \mathbf{w})$ for $\lambda > 0$;
4. $CE(\mathbf{x}, \lambda\mathbf{y}, \mathbf{w}) \geq CE(\mathbf{x}, \mathbf{y}, \mathbf{w})$ for $\lambda \geq 1$;
5. $CE(\mathbf{x}, \mathbf{y}, \lambda\mathbf{w}) = CE(\mathbf{x}, \mathbf{y}, \mathbf{w})$ for $\lambda > 0$.

Property 1 says that the measure of cost efficiency is bounded between zero and one. Property 2 states that a firm is cost efficient if, and only if, it achieves the minimum expenditure required to produce a given output vector, i.e. it uses a cost-minimising input vector. Property 3 says that CE ($\mathbf{x}$, $\mathbf{y}$, $\mathbf{w}$) is homogeneous of degree –1 in inputs, and Property 4 says that CE ($\mathbf{x}$, $\mathbf{y}$, $\mathbf{w}$) is nondecreasing in outputs. Property 5 states that CE ($\mathbf{x}$, $\mathbf{y}$, $\mathbf{w}$) is homogeneous of degree zero in input prices, indicating that the measure of cost efficiency depends only on relative input prices.

Technical efficiency is a necessary but not a sufficient condition for the achievement of cost efficiency. It may be the case that a technically efficient firm uses inappropriate mixes of inputs given the relative input prices it faces. This indicates the presence of allocative inefficiency in the firm.

3.4.3 Allocative Efficiency

A measure of input allocative efficiency can be introduced as the following function:

$$AE_I(\mathbf{x},\mathbf{y},\mathbf{w}) = CE(\mathbf{x},\mathbf{y},\mathbf{w}) / TE_I(\mathbf{x},\mathbf{y}) = C(\mathbf{y},\mathbf{w}) / \mathbf{w}^{\mathrm{T}}\theta\mathbf{x}. \tag{3.25}$$

The measure of input allocative efficiency satisfies the following properties (Kumbhakar and Lovell, 2000):

1. $0 < AE_I(\mathbf{x}, \mathbf{y}, \mathbf{w}) \leq 1$;
2. $AE_I(\mathbf{x}, \mathbf{y}, \mathbf{w}) = 1 \iff \exists\, \lambda \leq 1: \lambda\mathbf{x} = \mathbf{x}(\mathbf{y}, \mathbf{w})$;
3. $AE_I(\lambda\mathbf{x}, \mathbf{y}, \mathbf{w}) = AE_I(\mathbf{x}, \mathbf{y}, \mathbf{w})$ for $\lambda > 0$;
4. $AE_I(\mathbf{x}, \mathbf{y}, \lambda\mathbf{w}) = AE_I(\mathbf{x}, \mathbf{y}, \mathbf{w})$ for $\lambda > 0$.

Again, Property 1 says that the measure of input allocative efficiency is bounded between zero and one. Property 2 states that AE_I ($\mathbf{x}$, $\mathbf{y}$, $\mathbf{w}$) achieves its upper bound if, and only if, a radial contraction of a producer's input vector (due to technical inefficiency) results in a cost-minimising input vector. If no such contraction is possible, the producer is both allocatively and technically efficient, that is cost efficient. Properties 3 and 4 say that AE ($\mathbf{x}$, $\mathbf{y}$, $\mathbf{w}$) is homogeneous of degree 0 in inputs and input prices, respectively. These two properties imply that the measure of input allocative efficiency depends only on the relative input use and relative input prices.

3.4.4 Relationship between the Measures of Efficiency

From Eq.(3.25), the measure of cost efficiency decomposes to:

$$CE(\mathbf{x},\mathbf{y},\mathbf{w}) = TE_I(\mathbf{x},\mathbf{y}) \times AE_I(\mathbf{x},\mathbf{y},\mathbf{w}). \tag{3.26}$$

Cost efficiency has the property of multiplicative separability into input-allocative and technical efficiencies.[37] All three efficiency measures are bounded between 0 and 1. A firm is cost efficient if, and only if, it is both technically and allocatively efficient.

In Figure 3.1 a simple example of firms which use two inputs to produce a single output is illustrated. The set of input vectors $\mathbf{x} = (x_1, x_2)^T$ that are feasible for a chosen output y is represented by the input requirement set $L(y)$, whose boundary is given by an input isoquant *Isoq* $L(y)$. Technically efficient firms are presented by *Isoq* $L(y)$. If a given firm uses quantities of inputs defined by point $\mathbf{x}$ in Figure 3.1, it is technically inefficient since it lies above *Isoq* $L(y)$. The technical inefficiency of the firm is represented by the distance between points $\mathbf{x}$ and $\theta\mathbf{x}$, which is the amount by which all inputs could be proportionally reduced without a reduction in output. Technical efficiency θ can be expressed as the ratio between the distance from the origin to technically efficient input vector $\theta\mathbf{x}$ and the distance from the origin to input vector $\mathbf{x}$.

If the input price ratio, represented by the slope of isocost line $\mathbf{w}^T\mathbf{x}$, is known, then a cost efficient input combination can be identified. A firm that uses a cost-minimising input vector is presented by point $\mathbf{x}^*$, where isocost line $\mathbf{w}^T\mathbf{x}^*$ is a tangent to input isoquant *Isoq*$L(y)$. Thus, the minimum costs that can be achieved for the production of a given output y are $\mathbf{w}^T\mathbf{x}^*$. From Figure 3.1 we can see that the firm operating at $\theta\mathbf{x}$ is technically efficient but allocatively inefficient since it operates with higher costs (isocost line $\mathbf{w}^T\theta\mathbf{x}$ lies above the line $\mathbf{w}^T\mathbf{x}^*$). The distance between $\alpha\mathbf{x}$ and $\theta\mathbf{x}$ measures the allocative inefficiency of the firm. The allocative efficiency is defined as the ratio between the distance from the origin to $\alpha\mathbf{x}$ and the distance from the origin to $\theta\mathbf{x}$, whereas the total cost efficiency α can be calculated as the ratio between the distance from the origin to $\alpha\mathbf{x}$ and the distance from the origin to $\mathbf{x}$. It should be noted that, for a given output level y, the input combination $\alpha\mathbf{x}$ is not feasible since it lies outside the input requirement set $L(y)$. To reach the optimal input combination

[37] Separability may also be exploited in order to further decompose technical efficiency into scale, congestion, and 'pure' technical efficiency, as in Färe, Grosskopf, and Lovell (1985).

and thus become cost efficient, the firm would have to change its relative input use in the direction of increasing the use of input x_1 and decreasing the use of input x_2.

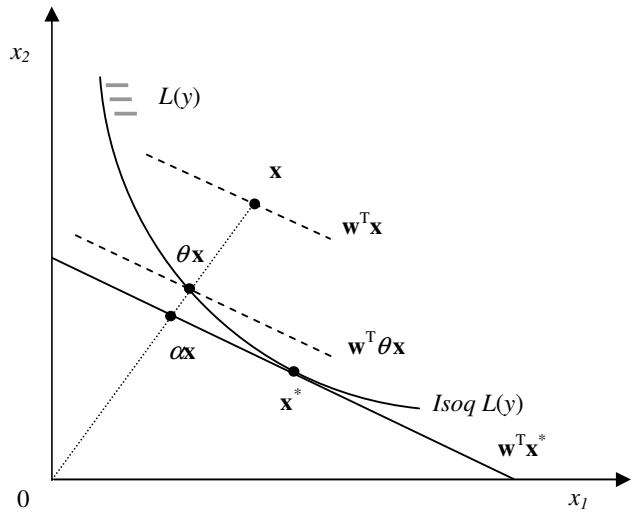

Figure 3.1: Decomposition of cost efficiency (Greene, 1997)

A final remark to be made refers to the cost minimizing behaviour. A question that can be raised in the presence of inefficiencies in the model is whether, by employing the cost function, one can truly arrive at the economic representation of the production possibility set. The theoretical cost function satisfying all the required properties laid down in Section 3.3 can only be derived in the presence of cost-minimising behaviour. If this is not the case, it is likely that some properties of the cost function, for example concavity in input prices, will not be satisfied by the estimated empirical cost function. There is thus no reason to assume that in the absence of cost-minimising behaviour the measured relationship between costs and outputs would represent technologically determined cost functions. One should keep in mind that in such cases the estimated empirical or frontier cost function cannot be viewed as the true cost function but rather as the 'behavioural' cost function (Evans, 1971, and Breyer, 1987).

3.5 Economies of Scale and Scale Efficiency

So far we have held output vector **y** fixed and discussed how to produce a given output vector with minimum input use or minimum cost. The former can be done by the contraction of input vector **x** to thus achieve technical efficiency, whereas the latter in addition requires

changing the input combination in order to achieve allocative efficiency as well. Besides the efficiency concepts described so far, an important role in cost-minimising behaviour is also played by economies or diseconomies of scale which help us define the optimal size of a firm. The optimal size is defined as the output level associated with the minimal average costs of production. We would thus like to establish whether for a proportional increase in vector of outputs one needs to increase the vector of inputs proportionally, more than proportionally or less than proportionally. Accordingly, the production technology can exhibit constant, nonincreasing or nondecreasing returns to scale as defined below (Mas-Colell et al., 1995):

- Nonincreasing returns to scale:
  if $(\mathbf{x},\mathbf{y}) \in T \Rightarrow (\alpha\mathbf{x},\alpha\mathbf{y}) \in T$ for $\alpha \in [0,1]$;
- Nondecreasing returns to scale:
  if $(\mathbf{x},\mathbf{y}) \in T \Rightarrow (\alpha\mathbf{x},\alpha\mathbf{y}) \in T$ for $\alpha \geq 1$;
- Constant returns to scale:
  if $(\mathbf{x},\mathbf{y}) \in T \Rightarrow (\alpha\mathbf{x},\alpha\mathbf{y}) \in T$ for any $\alpha \geq 0$.

The technology *T* exhibits nonincreasing returns to scale if any feasible input-output vector can be scaled down. This assumption is implied by convexity and the possibility of inaction. The production process exhibits nondecreasing returns to scale if any feasible input-output vector can be scaled up. The constant returns to scale property is just a conjunction of the first two properties.

A formal definition of economies of scale for the multi-product case states that the technology set *T* exhibits standard economies of scale at $(\mathbf{x},\mathbf{y}) \in T$ if, and only if, there is a $\delta > 1$ such that for all α with $1 < \alpha < \delta$, there is a $\gamma > \alpha$ with $(\alpha\mathbf{x},\gamma\mathbf{y}) \in T$.

Hence, there are economies of scale if a small proportional increase in the levels of all input factors can lead to a more than proportional increase in the levels of outputs produced. It should be noted that whenever $C(\mathbf{y}, \mathbf{w})$ is increasing in $\mathbf{y}$, this definition implies a decreasing ray average cost. The ray average cost (*RAC*) is the cost of an output vector of fixed proportions divided by a homogeneous measure of the size of the outputs (Baumol, 1977):

$$RAC = C(\gamma\mathbf{y},\mathbf{w})/\gamma. \tag{3.27}$$

The condition of decreasing ray average costs is then:

$$C(\gamma \mathbf{y}, \mathbf{w}) < \gamma C(\mathbf{y}, \mathbf{w}) \text{ for } \gamma > 1. \quad (3.28)$$

In the case of a single output Eq.(3.28) reduces to the familiar idea of declining average costs. It is evident that in this case one firm can produce a given output **y** less expensively than any group of firms. Hence, based on the presence of economies of scale the optimal size of the firm can be identified.[38] A natural monopoly allegedly minimises industry costs and is stable against entry if economies of scale are important and prevail over the full range of output. On the other hand, perfect competition can be viable only if firms' scale economies are exhausted at a level of output that is a small fraction of the market.

With more than one output, however, decreasing ray average costs only mean that an equiproportionate division of a monopolist's vector of outputs would increase industry costs. There may still be some division of the monopolist's outputs among several firms that decreases total industry costs (Panzar, Willig, 1977). Baumol (1977) argued that subadditivity should be the proper criterion for defining a natural monopoly since it implies that every output combination is always produced more cheaply by a single firm.[39]

Besides the standard definition of scale economies in a multi-output firm discussed above, there is a stronger definition according to which the technology set T exhibits economies of scale at $(\mathbf{x}, \mathbf{y}) \in T$ if there exists $r > 1$ and $\alpha > 1$ such that $(\alpha \mathbf{x}, \alpha^r \mathbf{y}) \in T$ for $1 \le \alpha \le \delta$.

This definition, first proposed by Panzar and Willig (1977), proves to be very useful since it allows the defining of a concept similar to the known concept of scale elasticity in the case of a single-output firm. This condition is local in that it is specific to a point (**x**, **y**) and only requires that $(\alpha \mathbf{x}, \alpha^r \mathbf{y}) \in T$ for α in an arbitrarily small neighbourhood. If the production process were homogeneous of degree $t > 1$, i.e. $(\mathbf{x}, \mathbf{y}) \in T$ would imply that $(\alpha \mathbf{x}, \alpha^t \mathbf{y}) \in T$, then the definition would hold for any α and with α invariant over T. In this definition, however, r may depend on the particular point (**x**, **y**). Scale economies according to the latter definition imply standard scale economies according to the former definition.

The properties of T were already provided in Section 3.1. We continue by further assuming that T is representable by a multi-output production transformation function, which is continuously differentiable in **x**, and continuously differentiable in y_m (m =1, …, M),

[38] More on the economies of scale, size and density concepts in the case of a single output firm will be given in Section 3.5.2.

[39] This discussion will be continued in Section 3.5.1.

for $y_m > 0$, at points (**x**, **y**) where **x** is cost efficient for **y**. This is a strong regularity condition on the smoothness of the technology which assures that isoquants have no corners. When this condition is satisfied, it is possible to define the degree of scale economies, *S*, in terms of multi-output technology. The degree of scale economies at $(\mathbf{x},\mathbf{y}) \in T$ is:

$$S = \sup\left\{r : \exists \delta > 1 \text{ such that } (\alpha\mathbf{x}, \alpha^r \mathbf{y}) \in T \text{ for } 1 \le \alpha \le \delta\right\}. \tag{3.29}$$

S is defined for all technologies since it is a local measure that is permitted to vary from point to point. *S* is positive on the assumption that inputs are always productive for all outputs. Economies of scale pertain if $S > 1$. Panzar and Willig (1977) showed that if the technology is homogeneous of degree *t*, then $S = t$ at all cost efficient points, i.e. at all $(\mathbf{x},\mathbf{y}) \in T$ where $\mathbf{w}^{\mathrm{T}}\mathbf{x} = C(\mathbf{y}, \mathbf{w})$. *S* is thus an indicator of the local degree of homogeneity of technology.

Further, *S* is a generalisation of the scale elasticity *E* of a scalar output production function. It is the standard result that *E* measures the ratio of average to marginal cost or, equivalently the ratio of cost to the revenue from marginal cost pricing when *E* is defined from the derivatives of the single-output differentiable production function. Panzar and Willig (1977) proved that at ($\mathbf{x}^*$, **y**), with $\mathbf{x}^* = \mathbf{x}$ (**y**, **w**) being cost-minimising input vector, *S* can be defined in a similar manner: $S = C(\mathbf{y},\mathbf{w}) \Big/ \sum_{m=1}^{M} \mathbf{y}_m C_m(\mathbf{y},\mathbf{w})$, where C_m(**y**, **w**) is a partial derivative with respect to y_m and stands for the marginal cost of *m*-th output produced. *S* thus measures the ratio of the production cost to the revenue that would result from marginal cost pricing in the case of multi-output production.[40] For a single-output production function, $S = E$.[41]

Taking into account the duality theory[42], we can obtain another representation of degree of scale economies, $\hat{S}$:

$$\hat{S} = \sup\left\{r : \exists \delta > 1 \text{ such that } C(\alpha y, w) \le \alpha^{1/r} C(y, w) \text{ for } 1 \le \alpha \le \delta\right\}. \tag{3.30}$$

[40] This also proves that in the presence of economies of scale, $S > 1$, marginal cost pricing is unprofitable since it does not allow a firm to cover all its costs.

[41] The single-output case will be examined in more detail in Section 3.5.2.

[42] According to modern duality theory, the theory of production can be developed treating the cost function as the primitive instead of the production set. That is, from observed behaviour in the form of costs, input prices and input demands, we are able to extract important characteristics of the technology. In McFadden's (1978) terminology, the cost function is a 'sufficient statistic' for the technology.

It can be shown that $S = \hat{S}$ (Panzar and Willig, 1977). The technology exhibits economies of scale, diseconomies of scale or locally constant returns to scale at ($\mathbf{x}^*$, $\mathbf{y}$) if and only if $\hat{S} > 1$, $\hat{S} < 1$ or $\hat{S} = 1$, respectively.

The relationship between scale economies and scale efficiency is discussed in Färe, Grosskopf and Lovell (1988). Scale efficiency is shown to correspond to constant returns to scale. In this case, the output level is associated with the minimal average cost of production. The source of input scale inefficiency can be either the production of an inefficiently small output vector in a region of increasing returns to scale or the production of an inefficiently large output vector in a region of decreasing returns to scale. If a firm chooses to operate at a production level where technology exhibits increasing or decreasing returns to scale it is regarded as scale inefficient since it can decrease its average costs by changing the production level to the point where constant returns to scale prevail.

3.5.1 Cost Subadditivity and Natural Monopoly

In essence, a natural monopoly arises if technology and demand are such that it is cheaper for one firm to serve the market than for several firms to do so. In such circumstances, competition is unfeasible and, hence, a monopoly seems to be 'natural'. A natural monopoly is usually associated with economies of scale. Economies of scale are present if a proportional increase in the levels of all inputs leads to a more than proportional increase in the levels of outputs produced (as defined in the previous section). In a single output case, this would lead to declining average costs meaning that one firm can produce a given output less expensively than any group of firms. In the multi-product case, however, economies of scale only imply that the monopolist's production would be cheaper than if the monopolist's vector of outputs were equiproportionately divided among any given number of firms, whereas this might not hold for an arbitrary division of outputs. Therefore, Baumol (1977) introduced the notion of subadditivity to define a natural monopoly in the multiple products case. Subadditivity of the cost function means that the cost of the sum of any N output vectors is less than the sum of the costs of producing them separately. More formally, a multi-product cost function $C(\mathbf{y}, \mathbf{w})$ is strictly and globally subadditive in the set of outputs in $M = 1, \ldots, m,$ if for any N output vectors $\mathbf{y}^1, \mathbf{y}^2, \ldots \mathbf{y}^N$ of the goods in M we have (Baumol, 1977):

$$C(\mathbf{y}^1 + \ldots + \mathbf{y}^N, \mathbf{w}) < C(\mathbf{y}^1, \mathbf{w}) + \ldots + C(\mathbf{y}^N, \mathbf{w}) . \tag{3.31}$$

This is a necessary and sufficient condition for a natural monopoly of any output combination in the industry producing (any and all) commodities in *M* because subadditivity means that it is always cheaper to have a single firm produce whatever combination of outputs is supplied to the market. It is possible that for some output vectors an industry will be a natural monopoly while for others it will not, in which case we have output-specific subadditivity.

The examination of a multi-product firm has led to additional developments in the area of economies of scope. Economies of scope are an extension of the notion of joint production and a particular case of the more general concept of subadditivity. Economies of scope are said to occur when it is possible for a single firm to produce two or more products more cheaply than it is possible to produce them with more than one firm. In the case of two products y_1 and y_2, economies of scope exist when

$$C(y_1, y_2; w) < C(y_1, 0; w) + C(0, y_2; w), \tag{3.32}$$

i.e., joint production is cheaper than separate production. Panzar and Willig (1981) provided a more rigorous definition of this and also showed that economies of scope exist if and only if the cost function with respect to the input shared by each output is subadditive.

3.5.2 Economies of Size, Output Density and Customer Density

A vast amount of the economic literature has been devoted to the effect of output on costs. We now turn to an investigation of this issue and confine our attention to a single output case. As already established, costs cannot decrease as output increases. This implies that marginal cost, defined as the partial derivative of the cost function with respect to output, is always nonnegative.

The elasticity of cost with respect to the output, or cost flexibility (Chambers, 1988), can be defined as follows:

$$\begin{aligned} \varepsilon_{\mathbf{y}}(\mathbf{y}, \mathbf{w}) &= \frac{\partial \ln C(\mathbf{y}, \mathbf{w})}{\partial \ln \mathbf{y}} \\ &= \frac{[\partial C(\mathbf{y}, \mathbf{w}) / \partial \mathbf{y}]^{\mathrm{T}} \mathbf{y}}{C(\mathbf{y}, \mathbf{w})}. \end{aligned} \tag{3.33}$$

The general definition in Eq.(3.33) applies to a multi-output case. In a single output case, Eq.(3.33) can be interpreted as the ratio of marginal cost divided by average cost:

$$\varepsilon_y(y,\mathbf{w}) = \frac{MC(y,\mathbf{w})}{AC(y,\mathbf{w})}. \tag{3.34}$$

The reciprocal of the elasticity of cost is referred to as the elasticity of size:

$$E_S = \varepsilon_y(y,\mathbf{w})^{-1}. \tag{3.35}$$

Here, a distinction has to be made between elasticity of scale and elasticity of size. Although the two concepts are closely related, they are only equal if the production technology satisfies certain additional requirements. Elasticity of size (E_S) measures the percentage increase in cost due to an increase in output. On the other hand, elasticity of scale (E) measures the percentage increase in output as a result of a proportional increase in all inputs:

$$E = \mathbf{i}^{\mathrm{T}} \frac{\partial \ln y}{\partial \ln \mathbf{x}}, \tag{3.36}$$

where $\mathbf{i}$ is a $K\times1$ vector of ones. Due to a possible reallocation of an input-minimising bundle (i.e., change in relative input use) caused by the output change, a one percent change in output need not be associated with a one percent change in all inputs. Thus, the two measures are generally not the same. These two measures only correspond in the case of a homothetic production function (Chambers, 1988). The cost function is consistent with the homotheticity of the production function whenever the cost function is separable, that is output is separable from input prices: $C(y,\mathbf{w}) = h(y)c(\mathbf{w})$. This also implies that elasticity of size is independent of input prices. Since any homogenous function is also homothetic, these results also apply to homogeneous functions.

If ε_y (y, $\mathbf{w}$) > 1 (or equivalently, E_S < 1) the firm exhibits diseconomies of size and smaller-sized operations (in terms of output produced) are more cost-effective in the sense that they are more scale efficient. If ε_y (y, $\mathbf{w}$) < 1, the firm exhibits economies of size (E_S > 1). Hence, larger-sized operations bring notable cost advantages. When ε_y (y, $\mathbf{w}$) = 1, the firm is characterised by constant returns to size (E_S = 1). In this case the cost increases with the

same proportion as the output so there are no gains or losses resulting from expanding or shrinking production. The firm is said to operate at the optimal level.[43]

In the case of network industries, the output typically possesses several dimensions. Besides output distributed, several output characteristics such as the number of customers, size of service area or length of network can influence the costs. According to Caves, Christensen and Tretheway (1984) and Roberts (1986), the inclusion of the number of customers and the size of service area in the cost function allows us to distinguish between economies of output density, economies of customer density and economies of size.

Assume that output vector **y** consists of the main output (Q) and two output characteristics, namely the number of customers (CU) and area size (AS). First, the economies of output density E_{OD} are defined in the following way:

$$E_{OD} = \frac{1}{\dfrac{\partial \ln C(Q, CU, AS, \mathbf{w})}{\partial \ln Q}}$$

$$= 1/\epsilon_Q \,, \tag{3.37}$$

where ϵ_Q is the elasticity of cost with respect to the output delivered (Q). Economies of output density measure the reaction of costs to an increase in output (Q), holding the number of customers and the size of the service area constant. It also follows that the customer density, defined as a ratio of the number of customers to the area size, is held constant.

Second, the economies of customer density E_{CD} are defined as follows:

$$E_{CD} = \frac{1}{\dfrac{\partial \ln C(Q, CU, AS, \mathbf{w})}{\partial \ln Q} + \dfrac{\partial \ln C(Q, CU, AS, \mathbf{w})}{\partial \ln CU}}$$

$$= 1/[\epsilon_Q + \epsilon_{CU}] \,, \tag{3.38}$$

[43] For instance, for a 'U-shaped' average-cost curve, average costs are minimised at the point where there are constant returns to size. Increasing returns to size are associated with decreasing average costs, whereas decreasing returns to size are associated with increasing average costs.

where ϵ_{CU} is elasticity of cost with respect to the number of customers (CU). Economies of customer density measure the reaction of costs due to a proportional increase in both the output and number of customers, holding the area size constant. In addition, it is assumed that, on average, new customers consume as much as the existing ones, i.e. output per customer is held fixed. This measure allows us to analyse an existing service area which is becoming more densely populated.

Finally, economies of size E_S are defined as:

$$E_S = \frac{1}{\frac{\partial \ln C(Q,CU,AS,\mathbf{w})}{\partial \ln Q} + \frac{\partial \ln C(Q,CU,AS,\mathbf{w})}{\partial \ln CU} + \frac{\partial \ln C(Q,CU,AS,\mathbf{w})}{\partial \ln AS}}$$

$$= 1/[\epsilon_Q + \epsilon_{CU} + \epsilon_{AS}] \,, \tag{3.39}$$

where ϵ_{AS} is the elasticity of cost with respect to the area size (AS). Alternatively, one could also consider the network length as a proxy for area size in Eq.(3.39). Economies of size measure the reaction of costs when the output, number of customers and area size increase proportionally. This measure becomes important when analysing whether or not it is beneficial to expand the size of the service area. This could be, for example, achieved by merging utilities. It is assumed that customer density and output per customer are held fixed.

It is said that there are economies, diseconomies or constant returns to size, customer density or output density if E_S, E_{CD} and E_{OD} are greater than unity, less than unity or equal to unity, respectively.

From the variable cost function one can also obtain the measure of economies of size as follows (Caves, Christensen and Swanson, 1981):

$$E_S = \frac{1 - \frac{\partial \ln VC(Q,CU,AS,\mathbf{w}_1,K)}{\partial \ln K}}{\frac{\partial \ln VC(\cdot)}{\partial \ln Q} + \frac{\partial \ln VC(\cdot)}{\partial \ln CU} + \frac{\partial \ln VC(\cdot)}{\partial \ln AS}}, \tag{3.40}$$

where K is the capital stock. Economies of output density and customer density can be obtained in a similar manner, i.e. by properly adjusting the numerator of Eq.(3.37) and Eq.(3.38), respectively. Measures obtained in this way refer to the long run.

By simply applying Eq.(3.37) and Eq.(3.38) to the variable cost function, short-run economies of density are obtained. Instead of short-run economies of density, some authors also speak of economies of utilisation (Caves, Christensen and Swanson, 1981). With respect to the economies of size, Garcia and Thomas (2001) point out that it is not interesting to consider the case where capital is not modified because merging several utilities into a single utility cannot be done without the consolidation of production and distribution facilities. In a similar way, if a utility wishes to expand its operation to cover a new area new investments are needed (e.g., expanding distribution network), which necessarily means expanding the capital stock, with this not being possible in the short run. Thus, in general one does not speak of short-run economies of size.

4 Functional Forms

In order to be able to estimate the cost function, we first have to specify a functional form. The decision on which functional form to choose for the empirical analysis is usually not straightforward since the true shape of the cost or production function is unknown. When choosing the functional form, one must keep the goal of the study firmly in mind. Within the context of the problem, the form should be as general as possible and impose the fewest possible *a priori* constraints or maintained hypotheses. Choosing a functional form limits the range the analysis can have. Once a general model is specified, classical statistical tests can only be conducted on the presumption that the general model is valid. Nevertheless, classical statistical theory is silent about the choice of functional form. Ideally, the theory suggests the form but many functional forms complying with the theory can be found. Choosing a functional form thus requires both a judgement and knowledge (Chambers, 1988).

The primary goal of an applied production analysis is to empirically measure economically relevant information that exhaustively characterises the behaviour of economic agents. For smooth technologies (i.e., those that are twice continuously differentiable), this includes the value of the function (the level of cost), the gradient of the function (the conditional factor demands) and the Hessian matrix (the conditional factor demand elasticities). One should try to find a form that is rich enough in parameters and can consequently estimate these effects independently and without imposing intrinsic restrictions or maintained hypotheses. Of special concern when analysing producers' behaviour are the maintained hypotheses on homogeneity, homotheticity, elasticity of substitution and concavity.

In the existing literature there is a wide variety of functional forms. The properties of the cost function established in the previous chapter will be used to determine possible advantages and disadvantages of applying a certain functional form when estimating the cost function. Some functional forms will be found to be too restrictive, imposing several restrictions upon the parameters of the cost function while, for the other, more flexible functional forms, we will have to verify whether all relevant properties of the cost function are satisfied. For example, Cobb-Douglas, CES and Leontief are more restrictive functional forms while others, like translog (or transcendental logarithmic), quadratic mean of order p and Generalised Leontief, are considered more flexible forms. Forms that can be either second-order differential or second-order numerical approximations are referred to as flexible

functional forms.[44] Flexible forms place no restrictions on the value of the function or its first or second derivatives at approximation point. In contrast, Cobb-Douglas is at best a first-order approximation.

It is also interesting to note that both the translog and quadratic mean of order *p* functional forms belong to a broader class of generalised quadratic forms, which is the class of locally flexible functional forms. Following Blackorby, Primont and Russel (1977), this class of functions can be expressed as:

$$h(\mathbf{z}) = \alpha + \sum_{i=1}^{n} \beta_i g_i(z_i) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij} g_i(z_i)\, g_j(z_j), \tag{4.1}$$

where $\beta_{ij} = \beta_{ji}$, and each $g_i(z_i)$ is a known twice continuously differentiable function of z_i.

Flexible forms are not without their shortcomings. As a rule, increased flexibility is associated with a greater need for information to adequately specify such relationships. Because of the increased number of parameters to be estimated, degrees of freedom are reduced and we might also end up with a problem of multicollinearity. Since reductions in maintained hypotheses come at a cost, added flexibility is not always desirable (Griffin, Montgomery, and Rister, 1987). The potential gains of choosing a more complex functional form must be balanced against difficulties involved in estimation, the structure imposed on the underlying production process and the ease with which parameter estimates can be interpreted. An 'ideal' functional form would minimise such a trade-off.

Further, flexible forms are very inflexible in representing separable technologies since a certain number of restrictions is required for separability to hold. Another limitation is their ability to approximate arbitrary technologies, which are local in nature. As approximations are not truly global, they cannot be exact for a wide range of observations.[45] Therefore, the

[44] Forms that can approximate any arbitrary twice continuously differentiable function are called second-order differential approximations, while forms that can be interpreted as a second-order, Taylor-series approximation to an arbitrary function are called second-order numerical approximations.

[45] The criticism of locally flexible functional forms has encouraged some authors to develop functional forms that are globally flexible. An example of such a functional form is the Augmented Fourier Form (Gallant, 1981) where no second-order restrictions are imposed anywhere in the domain. Augmented Fourier uses a trigonometric polynomial for the approximation and is quite a complex functional form. It requires the rescaling of variables so that they lie in the open interval $(0, 2\pi)$ and,

most likely contribution of the locally flexible forms lies in the fact that they apparently place far fewer restrictions prior to an estimation. They let measures like the elasticity of size and elasticity of substitution depend on data, i.e. they can vary across the sample and need not be parametric as they are for most of the more traditional forms (Chambers, 1988).

In what follows we provide the traditional and (locally) flexible functional forms most widely used in the empirical literature as well as criteria for selecting the most suitable functional form.

4.1 Different Functional Forms

4.1.1 The Cobb-Douglas Cost Function

The Cobb-Douglas cost function is empirically speaking the most widely exploited functional form. This functional form was introduced as a production function by Cobb and Douglas (1928). Since it is self-dual, the associated cost function has the same functional form. When firms are using K inputs to produce M outputs the Cobb-Douglas cost function can be written as follows:

$$C(\mathbf{y},\mathbf{w}) = \alpha \prod_{i=1}^{M} y_i^{\beta_i} \prod_{i=1}^{K} w_i^{\gamma_i} , \tag{4.2}$$

where C stands for cost, $\mathbf{y}$ is the vector of output(s) and output characteristics and $\mathbf{w}$ is the vector of input prices. In accordance with the properties of the cost function $\beta_i > 0$ and $\gamma_i > 0$, $\forall i$. If $n = K+M$ is the number of explanatory variables, the number of parameters to be estimated equals $1+n$. If we take the natural logarithm of Eq.(4.2), we obtain the following expression:

---

in addition to the constant term, linear and square terms, the inclusion of sine and cosine terms is required. It always has a larger number of parameters to be estimated than the locally flexible functional forms. In the case of relatively small samples, this form is not applicable due to the substantial loss of degrees of freedom. Therefore, the Fourier form will not be used in our empirical analysis and, accordingly, its presentation is considered to be beyond the scope of this work. Also, this functional form is not commonly used in production analysis.

$$\ln C(\mathbf{y},\mathbf{w}) = \ln \alpha + \sum_{i=1}^{M} \beta_i \ln y_i + \sum_{i=1}^{K} \gamma_i \ln w_i \,. \tag{4.3}$$

Eq.(4.3) is linear in parameters (and not in explanatory variables) and thus easy to estimate. The results are also easy to interpret.

The Cobb-Douglas cost function is linearly homogeneous in input prices if $\sum_k \gamma_k = 1$.[46] This restriction can be imposed in the estimation of the cost function by dividing the input prices and the cost by one of the input prices before taking logarithms (i.e. by the normalisation of input prices and costs). Another way is to impose a linear restriction in the estimation of Eq.(4.3). A method to test the homogeneity assumption is to estimate a restricted and unrestricted model and then to perform Wald test or likelihood ratio test.[47]

By applying Eq.(3.8) to Eq.(4.3) we find that a cost share for each input k equals γ_k (k = 1, …, K). Further, one of the restrictive properties of Cobb-Douglas is that it assumes all elasticities of substitution are equal to 1.

By applying Eq.(3.33) to Eq.(4.3) we obtain elasticities of cost with respect to each output y_m being equal to β_m (m = 1, …, M). Another restrictive assumption of the Cobb-Douglas cost function is constant economies of scale. Since the Cobb-Douglas functional form is homothetic, economies of scale coincide with economies of size. Economies of size are obtained as: $E_S = 1/\sum_m \beta_m$. If cost flexibility $\sum_m \beta_m < 1$, firms in the sample exhibit economies of scale and size. Economies of size that vary with output can be obtained by adding the square of the logarithmic output in Eq.(4.3).

4.1.2 The Translog Cost Function

The transcendental logarithmic function or translog was introduced by Christensen Jorgenson, and Lau (1971, 1973) and is one of the most commonly estimated flexible functional forms in the applied literature. The multi-output translog cost function is specified as follows:

[46] In this case the number of parameters to be estimated equals n.

[47] For example, see Greene (2000).

$$\begin{aligned}\ln C(\mathbf{y},\mathbf{w}) = \; & \ln\alpha + \sum_{i=1}^{M}\beta_i \ln y_i + \sum_{k=1}^{K}\gamma_k \ln w_k \\ & + \frac{1}{2}\sum_{i=1}^{M}\beta_{ii} \ln y_i \ln y_i + \sum_{i=1}^{M}\sum_{j=1}^{M}\beta_{ij} \ln y_i \ln y_j \\ & + \frac{1}{2}\sum_{k=1}^{K}\gamma_{kk} \ln w_k \ln w_k + \sum_{k=1}^{K}\sum_{l=1}^{K}\gamma_{kl} \ln w_k \ln w_l + \sum_{i=1}^{M}\sum_{k=1}^{K}\delta_{ik} \ln y_i \ln w_k ,\end{aligned} \tag{4.4}$$

where for the cross-product terms it holds that: $i \neq j$ and $k \neq l$. The expression in Eq.(4.4) is a logarithmic second-order Taylor approximation of an arbitrary function $\hat{C}(\mathbf{y},\mathbf{w})$ evaluated at $(\mathbf{y}^{\mathrm{T}}, \mathbf{w}^{\mathrm{T}}) = (1, 1, \ldots, 1)$.[48] Obviously, the means or medians of explanatory variables are considered to be much better representatives of a sample and therefore better expansion points. Usually, the median is considered the most appropriate approximation point since, as opposed to the mean, it is not affected by extreme values of explanatory variables.[49]

The translog cost function is a generalisation of the Cobb-Douglas functional form. If all second-order coefficients equal zero, the translog degenerates to the Cobb-Douglas form. The number of parameters to be estimated in the translog form is much larger compared to the Cobb-Douglas results. If there are n regressors, the number of parameters equals $\frac{1}{2}(n+1)(n+2)$.[50] As mentioned, having a large number of parameters to be estimated can result in a multicollinearity problem and a greater need for data.[51]

When estimating unknown technologies using flexible functional forms, it is not necessary to obtain a cost function, neither locally nor globally.[52] Therefore, Salvanes and Tjøtta (1998) proposed a procedure to calculate the consistency region, i.e. the region in which the

[48] Thus, when taking logarithms, ln$C$($\mathbf{y}$, $\mathbf{w}$) is evaluated around the point (0, 0, …, 0) which drops out of Eq.(4.4).

[49] This is achieved by the proper transformation of variables. All variables have to be divided by their median values before taking logarithms or, alternatively, from variables in a logarithmic form the respective logarithms of median values have to be subtracted.

[50] This is the sum of the n elements in the gradient vector, ½ n(n+1) distinctive elements of the Hessian and the constant term, which is the value of the function at the expansion point.

[51] To detect a possible multicollinearity problem, Maddala (2001) suggests examining t-values. Small t-values are a good indicator of multicollinearity. Due to multicollinearity, some coefficients might even end up having the wrong sign.

[52] Salvanes and Tjøtta (1998), for example, showed that Evans and Heckman's estimated cost function for the US Bell System (Evans and Heckman, 1984) is not a cost function since it was found to have a negative marginal cost in most of the test areas.

required regularity conditions are met, for a multi-output translog cost function. These regularity conditions consist of already known properties of non-negative costs, non-negative marginal costs, homogeneity in input prices and monotonicity and concavity of input prices.

Usually, the translog model is restricted to be (globally) linearly homogeneous in prices by:[53]

$$\sum_{i=1}^{K}\gamma_i = 1, \qquad \sum_{i=1}^{K}\gamma_{ij} = 0, \forall j, \qquad \text{and} \qquad \sum_{i=1}^{K}\delta_{ij} = 0, \forall j. \tag{4.5}$$

Symmetry conditions imply: $\beta_{ij} = \beta_{ji}$, $\gamma_{ij} = \gamma_{ji}$, $\delta_{ij} = \delta_{ji}$, $\forall i, j$. In the case of a twice continuously differentiable function they are automatically satisfied. The regularity condition that the estimated total cost is positive is met since $\exp(\ln C(\mathbf{y},\mathbf{w}))$ is strictly positive for all ($\mathbf{y}$, $\mathbf{w}$). The next condition is that marginal cost with respect to output is non-negative which holds if and only if:

$$\begin{aligned}\frac{\partial \ln C(\mathbf{y},\mathbf{w})}{\partial \ln y_i} &\equiv \frac{\partial C(\mathbf{y},\mathbf{w})}{\partial y_i}\frac{y_i}{C(\mathbf{y},\mathbf{w})} \equiv \epsilon_{yi} \\ &= \beta_i + \sum_{j=1}^{M}\beta_{ij}\ln y_j + \sum_{j=1}^{K}\delta_{ij}\ln w_j \quad \geq 0.\end{aligned} \tag{4.6}$$

Economies of output density, customer density and economies of size can then be estimated using Eq.(3.37), (3.38) and (3.39), respectively. As opposed to the Cobb-Douglas functional form, these measures vary with output and/or output characteristics, which is evident from Eq.(4.6).

By applying Shephard's lemma to Eq.(4.4) we can obtain the cost shares of each input. In addition, Shephard's lemma allows us to verify whether cost is nondecreasing in input prices, which holds if and only if:

[53] When estimating the translog, linear homogeneity is imposed in the same way as described for the Cobb-Douglas functional form. When linear homogeneity is imposed, the number of parameters to be estimated equals ½ M (M+1) + ½ K(K+1) + MK, where M and K are the number of outputs and inputs, respectively.

$$\begin{aligned}\frac{\partial \ln C(\mathbf{y},\mathbf{w})}{\partial \ln w_i} &\equiv \frac{\partial C(\mathbf{y},\mathbf{w})}{\partial w_i}\frac{w_i}{C(\mathbf{y},\mathbf{w})} \equiv S_i \\ &= \gamma_i + \sum_{j=1}^{K}\gamma_{ij}\ln w_j + \sum_{j=1}^{M}\delta_{ij}\ln y_j \quad \geq 0.\end{aligned} \tag{4.7}$$

Since linear homogeneity is imposed, the input cost shares S_i sum up to unity.[54] The last regularity condition is that the estimated cost function is concave in input prices. This corresponds to the Hessian matrix with respect to input prices being negative semidefinite. Following Diewert and Wales (1987), this Hessian is negative semidefinite if and only if

$$\Gamma(\mathbf{y},\mathbf{w}) = \begin{bmatrix} \gamma_{11} - S_1(1-S_1) & \gamma_{12} + S_1S_2 & \cdots & \gamma_{1K}S_1S_K \\ \gamma_{12} + S_1S_2 & \gamma_{22} - S_2(1-S_2) & \cdots & \gamma_{2K}S_2S_K \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{1K} + S_1S_K & \gamma_{2K} + S_2S_K & \cdots & \gamma_{KK} - S_K(1-S_K) \end{bmatrix} \tag{4.8}$$

is negative semidefinite. It is worth noting that the elements of the matrix $\Gamma(\mathbf{y}, \mathbf{w})$ depend on $(\mathbf{y}, \mathbf{w})$ since input shares S_i depend on $(\mathbf{y}, \mathbf{w})$. Imposing global concavity in input prices destroys the flexibility of the translog cost function. Thus, the concavity requirement is not imposed and has to be tested after the estimation of the cost function.[55]

If the translog cost function is to be consistent with a homothetic technology, the output must be separable from input prices. Taking $\exp(\ln C(\mathbf{y},\mathbf{w}))$ and using Eq.(4.4) it can be established that, if the function is to be globally homothetic, it must be true that: $\delta_{ij} = 0, \forall i, j$.

Since the translog functional form is a local approximation, the estimation results are reliable close to the approximation point. Observations far from this point may lead to the wrong conclusions. White (1980) further demonstrated that Ordinary Least Squares estimators of Taylor-series expansions are not reliable indicators of the parameter vector of the true

[54] The estimation of (4.4) together with the share equations requires a modification of Zellner's (1962) Seemingly Unrelated Regressions (SUR) method. Since the share equations add up to one, the covariance matrix of disturbances is singular. Thus, one share equation has to be dropped in the estimation procedure.

[55] As an alternative, Ryan and Wales (2000) propose a method for imposing concavity locally while at the same time maintaining flexibility in the case of the translog and generalised Leontief forms.

expansion of a known function. This is due to the fact that the Least Squares (LS) method weights all observations equally, while for the derivation of Taylor series only the approximation point is the point of interest. Parameters of the translog function estimated by the LS model are therefore almost always biased. As a consequence, predictive properties of locally flexible functional forms have been found to be satisfactory (for large samples), but inferences involving single parameter estimates are not reliable.

4.1.3 The Generalised Leontief Cost Function

The generalised Leontief functional form for a single output case was proposed by Diewert (1971) and has the following form:[56]

$$C(y,\mathbf{w}) = h(y)\sum_{i=1}^{K}\sum_{i=1}^{K}\gamma_{ij}\left(w_i w_j\right)^{\frac{1}{2}}, \tag{4.9}$$

where $h(y)$ is a continuous, monotonically increasing function. To satisfy the symmetry requirement, $\gamma_{ij} = \gamma_{ji} \geq 0$. It is often assumed that $h(y) = y$, which imposes constant economies of scale.

Since this form is not locally flexible when $h(y)$ is not known, Diewart and Wales (1987) suggested a locally flexible version of a generalised Leontief, which can be expressed as:

$$C(y,\mathbf{w}) = y\sum_{i=1}^{K}\sum_{i=1}^{K}\gamma_{ij}\left(w_i w_j\right)^{\frac{1}{2}} + \sum_{i=1}^{K}\gamma_i w_i + \beta_{yy}\left(\sum_{i=1}^{K}\delta_i w_i\right)y^2. \tag{4.10}$$

Similarly to the translog case, global concavity in input prices is not imposed in the estimation since this would destroy flexibility. The concavity has to be tested after the estimation.

Since this functional form is specified only for a single output case, Hall (1973) proposed a hybrid Diewert multi-product cost function which has the following expression:[57]

---

[56] Eq.(4.9) is a generalisation of Leontief in the sense that when setting $\gamma_{ij} = 0$ ($\forall i \neq j$) and $h(y) = y$ and applying Shephard's lemma, the cost-minimising input vector is given by: $x_i(y, \mathbf{w}) = \gamma_{ii}\, y$, which corresponds to Leontief production function.

$$C(\mathbf{y},\mathbf{w}) = \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{K}\alpha_{ijkl}\left(y_i y_j w_k w_l\right)^{\frac{1}{2}}. \tag{4.11}$$

Again, symmetry requires: $\alpha_{ijkl} = \alpha_{ijlk} = \alpha_{jikl} = \alpha_{jilk}$. The appeal of this functional form is the fact that it is linearly homogeneous in input prices without the imposition of any additional linear restrictions as in the translog case. Separability of the cost function in Eq.(4.11) is imposed by setting $\alpha_{ijkl} = \beta_{ij}\,\gamma_{kl}$, which results in:

$$C(\mathbf{y},\mathbf{w}) = \sum_{i=1}^{M}\sum_{j=1}^{M}\beta_{ij}\left(y_i y_j\right)^{\frac{1}{2}}\sum_{k=1}^{K}\sum_{l=1}^{K}\gamma_{kl}\left(w_k w_l\right)^{\frac{1}{2}}. \tag{4.12}$$

Nevertheless, the hybrid Diewert multi-product cost function cannot be classified as a flexible functional form since not all first- and second-order effects are unrestricted. Therefore, it cannot be viewed as an approximation of an arbitrary functional form. For example, although the hybrid Diewert multi-product cost function contains no *a priori* restrictions on elasticities of substitution among factor inputs, it imposes constant returns to scale. Generalising this form to permit flexibility in scale economies necessitates a large increase in the number of parameters (Caves, Christensen and Tretheway, 1980).

4.1.4 The Quadratic Cost Function

A third flexible form which might be used to represent a multi-output cost function is the quadratic functional form suggested by Lau (1974):[58]

[57] The number of parameters of this functional form equals $M\,(M+1)\,K\,(K+1)/4$, where M and K stand for the number of outputs and inputs. This exceeds the number of parameters in translog form, except when there are only two inputs and outputs.

[58] The number of parameters to be estimated equals the translog form. When the translog is restricted to be linearly homogeneous in prices, the quadratic form has $M+K+1$ more parameters than the translog, where M is the number of outputs and K is the number of inputs.

$$C(\mathbf{y},\mathbf{w}) = \alpha + \sum_{i=1}^{M}\beta_i y_i + \sum_{i=1}^{K}\gamma_i w_i + \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\beta_{ij} y_i y_j + \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K}\gamma_{ij} w_i w_j + \sum_{i=1}^{M}\sum_{j=1}^{K}\delta_{ij} y_i w_j . \tag{4.13}$$

Compared to the translog, variables in the quadratic functional form are not in logarithms and can therefore take on zero values as well. However, the quadratic cost function has one serious flaw. It is not linearly homogeneous in input prices nor can homogeneity be imposed by parametric restrictions without sacrificing the flexibility of the form (Caves, Christensen and Tretheway, 1980). Thus, the quadratic form is not an attractive form of the multi-product cost function and will be disregarded in what follows.

4.1.5 The Quadratic Mean of Order p

Besides the translog, Chambers (1988) also considers another cost function which is obtained as a second-order, Taylor-series approximation. Taking a second-order, Taylor-series approximation of the transformation of function $\hat{C}(\mathbf{y},\mathbf{w})$, $\hat{C}(\mathbf{y},\mathbf{w})^p$, in terms of $(\mathbf{y}^T,\mathbf{w}^T)^{p/2}$ in the neighbourhood of the null vector results in a quadratic mean of order p:[59]

$$C(\mathbf{y},\mathbf{w})^p = \alpha + \sum_{i=1}^{M}\beta_i y_i^{p/2} + \sum_{i=1}^{K}\gamma_i w_i^{p/2} + \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\beta_{ij} y_i^{p/2} y_j^{p/2} + \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K}\gamma_{ij} w_i^{p/2} w_j^{p/2} + \sum_{i=1}^{M}\sum_{j=1}^{K}\delta_{ij} y_i^{p/2} w_j^{p/2} . \tag{4.14}$$

Again, symmetry is guaranteed by $\beta_{ij} = \beta_{ji}$, $\gamma_{ij} = \gamma_{ji}$, $\delta_{ij} = \delta_{ji}$, $\forall i, j$. To guarantee the monotonicity and positivity of the cost function, it is necessary that none of the coefficients in Eq.(4.14) is negative. Linear homogeneity in input prices is once again not satisfied and we will therefore not consider this functional form as a suitable candidate for estimating the cost function.

[59] If this approximation is not taken in the neighbourhood of (0, 0, …, 0), explanatory variables have to be suitably normalised. Usually this is done by subtracting the mean or median value from each variable.

It is also interesting to note that both the translog and quadratic mean of order *p* functional forms belong to a broader class of generalised quadratic forms, which is the class of locally flexible functional forms. At the beginning of this section we already pointed out the main features of locally flexible functional forms and also provided some critical notes.

4.1.6 The Generalised Translog Multi-product Cost Function

One of the desirable characteristics of a multi-product cost function is that it permits a value of zero for one or more outputs. The quadratic cost function and generalised Leontief permit zero output values. However, in the translog functional form all of the outputs enter in a logarithmic form and therefore the translog has no finite representation if any output has a zero value. This flaw of the translog cannot be neglected since firms in a multi-product industry might only produce a subset of feasible outputs.

Nevertheless, the translog functional form can be generalised to permit zero output levels. It seems natural to retain the log metric for input prices and total cost, but for outputs to choose a metric that is well-defined for zero values. A metric is available that not only permits zero values but also contains the natural logarithm metric as a limiting case. This metric was proposed by Box and Cox (1964):

$$y_i^{(\pi)} = \begin{cases} (y_i^{\pi} - 1)/\pi & \text{for} \quad \pi \neq 0 \\ \ln y_i & \text{for} \quad \pi = 0. \end{cases} \tag{4.15}$$

Provided that π is strictly positive, the Box-Cox transformation is well-defined for zero output levels (it equals $-1/\pi$). The natural log transformation is a limiting case of the Box-Cox transformation.[60] The generalised translog multi-product cost function proposed by Caves, Christensen and Tretheway (1980) can thus be written in the following way:

$$\begin{aligned} \ln C(\mathbf{y}, \mathbf{w}) = \; & \ln \alpha + \sum_{i=1}^{M} \beta_i y_i^{(\pi)} + \sum_{i=1}^{K} \gamma_i \ln w_i \\ & + \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \beta_{ij} y_i^{(\pi)} y_j^{(\pi)} + \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} \gamma_{ij} \ln w_i \ln w_j + \sum_{i=1}^{M} \sum_{j=1}^{K} \delta_{ij} y_i^{(\pi)} \ln w_j . \end{aligned} \tag{4.16}$$

[60] By applying l'Hôpital's rule we obtain: $\lim_{\pi \to 0} \left(y_i^{\pi} - 1 \right) / \pi = \ln y_i$.

By imposing $\pi = 0$ we obtain the translog cost function specified in Eq.(4.4). One can also test the standard translog against the generalised translog specification. The requirement of linear homogeneity in input prices is met by imposing the same restrictions as for the translog. Compared to the standard translog functional form the generalised translog has one more parameter to estimate.[61] A further generalisation of Eq.(4.16) is presented in the following section.

4.1.7 The General Box-Cox Model

It is interesting to note that the generalised translog, as well as some other widely used alternative cost functions, is nested within the general Box-Cox model (1964):

$$C^{(\phi)}(\mathbf{y},\mathbf{w}) = \left\{ \exp\left[\left(\alpha + \sum_{i=1}^{M} \beta_i y_i^{(\pi)} + \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \beta_{ij} y_i^{(\pi)} y_j^{(\pi)} + \sum_{i=1}^{M} \sum_{j=1}^{K} \delta_{ij} y_i^{(\pi)} \ln w_j \right)^{(\tau)} \right] \right.$$
$$\left. \cdot \exp\left[\sum_{i=1}^{K} \gamma_i \ln w_i + \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} \gamma_{ij} \ln w_i \ln w_j \right] \right\}^{(\phi)}, \qquad (4.17)$$

where the superscripts in parentheses, ϕ, π and τ, represent Box-Cox transformations. The generalised translog is obtained by setting $\phi = 0$ and $\tau = 1$, while the standard translog requires a further restriction, $\pi = 0$.

Pulley and Braunstein (1992) further introduced the composite cost function for multi-product firms by setting $\pi = 1$ and $\tau = 0$ in Eq.(4.17). Moreover, a separable quadratic function can be obtained by adding another restriction to the composite cost function: $\delta_{ij} = 0$. The composite cost function is a combination of the log-quadratic input price structure of the translog form with a quadratic structure for outputs. It allows us to measure economies of scope, output-specific economies of scale and subadditivity even in the case of some zero output values.

Since the general Box-Cox model is nonlinear in parameters, it has to be estimated with a nonlinear maximum likelihood technique. For the functional forms that are nested within the general Box-Cox model, the relative statistical fitness can be determined using the standard

[61] One could further generalise the generalised translog by specifying a distinct π for each output or by replacing $y_i^{(\pi)}$ with $(y_i + \psi)^{(\pi)}$. However, these generalisations needlessly complicate the estimation.

likelihood ratio (LR) test. Specifically, each functional form can be tested against the general Box-Cox model.

4.1.8 The Hedonic vs. General Cost Function Specification

Most econometric studies appearing up until the mid-1970s ignored the heterogeneity of outputs and estimated the cost function with one or the two output measures. In general, the use of an aggregate output will yield unbiased empirical results only if the subcomponents of the aggregate vary in the same proportion (Panzar and Willig, 1977). This condition is not likely to be met in most situations. According to Oum and Tretheway (1989), recent approaches to incorporate the heterogeneity of outputs may be classified roughly in the following categories: (i) attempts to use disaggregate outputs; (ii) use of multiple aggregate outputs; and (iii) the use of single or multiple aggregate outputs and attribute or quality variables describing outputs. Recognising the impossibility of complete output disaggregation, an increasing number of econometric studies of cost functions have begun incorporating quality or attribute variables to describe the outputs. These variables are introduced in order to correct for differences in output mix between firms or from one time period to another.

Oum and Tretheway (1989) distinguished two approaches in the way output attributes are incorporated within the cost function. The two alternative approaches are a hedonic and a general specification of the cost function. The hedonic approach specifies the cost function in the following form:

$$C = C\left(\phi_1(y_1, \mathbf{q_1}), \ldots, \phi_M(y_M, \mathbf{q_M}), \mathbf{w}, t\right), \tag{4.18}$$

where y_m is the m-th output (m = 1, …, M), $\mathbf{q_m}$ is a $1 \times L_m$ vector of output attributes (l_m = 1,… L_m) for m-th output, and the time variable t is included to capture the effect of technical change. Eq.(4.18) is a hedonic cost function, while the $\phi_m(\cdot)$ imbedded in the cost function are referred to as hedonic output aggregator functions. The hedonic output specification attempts to adjust the observed outputs for the variation in their quality attributes.[62] This aggregation requires the separability of the arguments in each hedonic

[62] Microeconomic foundations for conducting the formal analysis and measurement of quality attributes were provided by the work of Lancaster (1966) and Rosen (1974). Once the existence of quality adjusted price and volume indices is accepted, the standard results of neoclassical theory also hold in quality-adjusted price and quantity spaces.

aggregator from all other arguments in the hedonic cost function.[63] This is a restrictive assumption and may not be empirically valid for some problems. Restricting the hedonic functional form to a separable structure is certainly a limitation of the hedonic specification of outputs in the cost function.

Usually, the hedonic cost function to be estimated is specified in the translog form. In this case, output aggregator functions ϕ_m take the Cobb-Douglas functional form which is another limitation of the hedonic specification.[64] Accordingly, Eq.(4.18) can be written as:

$$\begin{aligned}\ln C = \ & \ln\alpha + \sum_{i=1}^{M}\beta_i \ln\phi_i(y_i,\mathbf{q_i}) + \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\beta_{ij}\ln\phi_i(y_i,\mathbf{q_i})\ln\phi_j(y_j,\mathbf{q_j}) \\ & + \sum_{i=1}^{K}\gamma_i \ln w_i + \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K}\gamma_{ij}\ln w_i \ln w_j + \sum_{i=1}^{M}\sum_{j=1}^{K}\delta_{ij}\ln\phi_i(y_i,\mathbf{q_i})\ln w_j .\end{aligned} \tag{4.19}$$

The second approach is to include the output attributes in the cost function in a general manner without imposing any restriction on the functional structure, which can be expressed as follows:

$$C = C(y_1,\ldots,y_M,\mathbf{q_1},\ldots,\mathbf{q_M},\mathbf{w},t)\,, \tag{4.20}$$

where all variables are defined in the same way as above. Eq.(4.20) is a general specification of the model, without any restrictions imposed. Similarly, the translog form of Eq.(4.20) results in:

$$\begin{aligned}\ln C = \ & \ln a + \sum_{i=1}^{M} b_i \ln y_i + \tfrac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} b_{ij}\ln y_i \ln y_j \\ & + \sum_{k=1}^{K} c_k \ln w_k + \tfrac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{K} c_{kj}\ln w_k \ln w_j + \sum_{i=1}^{M}\sum_{k=1}^{K} d_{ik}\ln y_i \ln w_k\end{aligned}$$

[63] The ratios of the marginal effects of y_m and q_{ml} on the cost are assumed to be independent of the level of any other arguments of the hedonic cost function which are not included in ϕ_m.

[64] Formal arguments can be found in Denny and Fuss (1977). It is shown that the translog hedonic cost function must be either a Cobb-Douglas function of translog aggregates or a translog function of Cobb-Douglas aggregates. In the latter case, the Cobb-Douglas form in turn implies homotheticity of the hedonic aggregator as well as the separability of output y_m and attributes q_{ml} from other arguments in the cost function. Thus, besides separability, an additional limitation is imposed in the case of the translog hedonic cost function.

$$+\sum_{m=1}^{M}\sum_{l=1}^{L_m} f_{ml}\ln q_{ml}+\tfrac{1}{2}\sum_{m=1}^{M}\sum_{l=1}^{L_m}\sum_{i=1}^{M}\sum_{j=1}^{L_m} f_{ijml}\ln q_{ml}\ln q_{ij}$$
$$+\sum_{i=1}^{M}\sum_{m=1}^{M}\sum_{l=1}^{L_m} g_{iml}\ln y_i \ln q_{lm}+\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{l=1}^{L_m} h_{iml}\ln w_k \ln q_{lm}. \qquad (4.21)$$

In Oum and Tretheeay (1989) the hedonic translog cost function is shown to be nested within the general translog specification in Eq.(4.21) through a set of nonlinear constraints on the parameters of the general translog function. Thus, the hedonic specification can be empirically tested, given sufficient degrees of freedom to estimate the general translog function.

Blackorby, Primont, and Russel (1977) showed that once separability is imposed as in Eq(4.19), the translog specification of the cost function is no longer capable of providing a second-order approximation to any unknown arbitrary separable cost function. The translog function with hedonic output specifications must then be interpreted as an exact form of the cost function, not as an approximation. This is another serious limitation of the hedonic specification. Further, when we embed output aggregator functions into the hedonic translog specification the parameters are under-identified since there are more parameters to be estimated than regressors. To be able to estimate the translog hedonic cost function, a suitable normalisation of the hedonic parameters has to be imposed.[65] In the case of the Cobb-Douglas output aggregator function we can write:

$$\ln \phi_i(y_i, \mathbf{q_i}) = \eta_{i0}\ln y_i + \sum_{l=1}^{L_i}\eta_{il}\ln q_{il}\,. \qquad (4.22)$$

Substituting Eq.(4.22) in Eq.(4.19) we arrive at the final form of the cost function. To solve the identification problem, researchers typically impose a normalisation of the hedonic parameters by constraining $\eta_{i0} = 1$.[66] These M restrictions seem theoretically sensible in that they make the hedonic (quality adjusted) outputs, ϕ_m, linearly homogeneous in observed

[65] For details, see Oum and Tretheway (1989).

[66] See, for example, Spady and Friedlaender (1978) and Feigenbaum and Teeples (1983). It may be as well noted that the latter study employed a translog hedonic cost function with the translog form of the cost function and the Cobb-Douglas functional form of the output aggregator function. The former study incorrectly used the translog functional form for both the cost function and output aggregator function. The resulting function in this case is not a translog hedonic cost function. Recall that either cost function or output aggregator should take on the Cobb-Douglas functional form.

outputs, y_m. By substituting Eq.(4.22) in Eq.(4.19) we in fact obtain a general translog function. Hence, the usual procedure is to estimate the general translog function with the appropriate (non-linear) restrictions imposed on its parameters. The resulting system is nonlinear in parameters and has to be estimated with a nonlinear maximum likelihood technique.

Up to this point it has been shown that the hedonic specification of the cost function is quite restrictive and it therefore seems to be unattractive for empirical research. The estimation of a multi-product cost function, however, requires a large data set and the general translog form in Eq.(4.21) needs many more parameters to be estimated than the translog hedonic function.[67] The general form is far less restrictive but it consumes valuable degrees of freedom. As a result, this reduces the number of attribute variables that can be introduced in the cost function. There is again a trade-off and one must choose between a richer specification in terms of the number of attribute variables (hedonic approach) and a richer specification in terms of fewer arbitrary restrictions on the structure of costs.

4.2 Selection Criteria for a Functional Form

The researcher is never in a position to know the true functional form so the problem is to choose the best form for a given task. With the growing number of functional forms available, the model builder's task is becoming ever more complicated. A comparison of the different functional forms requires some *a priori* selection criteria which should refer to mathematical, statistical and economic properties and are useful for formalising the selection of the functional form during the model-building process. By combining the criteria in Lau (1986) with those in Griffin, Montgomery and Rister (1987) the following conditions for selecting a functional form are defined:

- theoretical consistency and domain of applicability;
- flexibility vs. maintained hypotheses;
- statistical estimation; and
- general conformity to data.

[67] For the case of K input prices, M outputs and L attribute variables ($L = \Sigma L_m$), the number of parameters to be estimated in the general translog form equals $[K(K+1)/2 + M(M+1)/2 + L(L+1)/2 + KL + KM + ML]$, while the hedonic translog specification requires only $[K(K+1)/2 + M(M+1)/2 + KM + L]$ parameters.

To meet the theoretical consistency condition, a selected functional form has to satisfy certain properties indicated by economic theory. The cost function thus has to be nondecreasing in output, linearly homogeneous in input prices and nondecreasing and concave in input prices. Linear homogeneity in input prices is usually imposed prior to the estimation. All functional forms except for the quadratic and quadradic mean of order one meet this requirement. Concavity in input prices is normally not imposed *ex ante* since this would destroy the flexibility of the flexible functional forms and would thus have to be tested *ex post*. This is done by evaluating the Hessian at any point of interest. The consistency conditions are not necessarily met in all observed data points so another issue to be addressed is that of the domain of applicability. Essentially, we are interested in identifying regions where the required regularity conditions are met. As approximations by examined flexible forms are not global, the regularity conditions may not be valid for a wide range of observations. Once again, imposing these conditions globally would destroy the flexibility since additional restrictions on the parameters would have to be imposed. One can test after the estimation for which data points the regularity conditions are met.

The second criterion deals with flexibility as opposed to maintained hypotheses. Concerns regarding maintained hypotheses can be used to assess the appropriateness of the functional form. If the maintained hypotheses implied by a certain form are acceptable or even useful, then the function might be considered appropriate. In the absence of a strong theoretical or empirical basis for adopting a given maintained hypothesis, a functional form which is unrestrictive with respect to this hypothesis may be considered appropriate. As already discussed, locally flexible functional forms place far fewer restrictions before the estimation than the traditional functional forms like Cobb-Douglas, Leontief or CES. Usually, flexible functional forms subsume one or more traditional forms as special cases. A detailed list of this can be found in Griffin, Montgomery and Rister (1987). The authors also provide all relevant properties of the functional forms most used in the production analysis.

The criterion for statistical estimation encompasses several aspects. First, unknown parameters should be easy to estimate from the data. This holds for functional forms that are linear in parameters. These functions permit parameter estimation by linear least squares methods. Sometimes linearity in parameters is achieved after applying a known transformation to the function. Out of all the functions discussed, the general Box-Cox model, the composite cost function, the generalised translog and the translog hedonic cost function do not meet this condition. Second, the functional form should be expressed in an explicitly closed form, which holds for all the functions discussed here. Further, the choice of functional form is also based on data availability and the number of variables we wish to include in the function. Most functions require a geometrically growing number of

parameters to be estimated as the number of explanatory variables is increased. This is primarily due to the large number of interactions specified among explanatory variables. Fuss, McFadden and Mundlak (1978) also pointed out the presence of multicollinearity as an increasing number of variables is included in the model.[68] Moreover, the researcher might take into the account the expense of estimating a large number of parameters in terms of the loss of degrees of freedom. This concern becomes especially important when we deal with a relatively small number of observations.

The last criterion is related to the general conformity of data, according to which a functional form should be consistent with empirical facts. This criterion is related to a specific dataset so the findings are typically not general.

To summarise, a functional form may be appropriate because of its flexibility, the correspondence of maintained hypotheses with the theory, the possibility and ease of statistical estimation, the possibility and ease of application, general conformity to the data, or a combination of these criteria. None of these criteria guarantee that the true relationship will be discovered nor do any allow a perfectly objective choice to be made. A subjective judgement is a necessary aspect of choice regarding the functional form. Having selected two or more estimable functional forms with plausible theoretical and applicative properties, the researcher may wish to base their final decision on statistical criteria. Such criteria clearly entail data-specific considerations. In the empirical part of the thesis in Chapter 6 these considerations will be taken into account when choosing the most appropriate functional form for estimating the cost function for Slovenian water distribution utilities.

[68] If multicollinearity is high, the variance of the parameter estimates is increased such that it may be impossible to determine how much variation in the endogenous variable is explained by different exogenous variables. Due to the higher parameter variances, some variables might also turn out to have an insignificant influence on the cost.

5 Parametric Methods for Estimating Cost Inefficiency

After having introduced efficiency concepts in Chapter 3 and analysed in detail the properties of the (frontier) cost function and alternative functional forms in Chapter 4, we now turn to the problem of estimating cost (in)efficiency. The estimation of frontier cost functions can be viewed as an econometric problem of making the empirical implementation consistent with the theoretical proposition that no observed agent can exceed the ideal. Cost inefficiency scores can then be obtained from the estimated cost frontier function as the deviation from the optimal point on the cost frontier. In the literature there are many different methods to estimate the cost frontier. Based on whether they allow for a stochastic error or not, parametric frontier methods are divided into deterministic and stochastic methods. Deterministic frontier can be estimated by the Corrected Ordinary Least Squares (COLS) method, while the class of stochastic frontier methods offers a much broader choice of methods. Depending on the nature of the data set the SFA methods can be classified into cross-sectional and panel data models. Panel data models further allow for many different possibilities for estimating the cost frontier and corresponding inefficiencies. Since panel data contain observations of companies over a certain time period firm-specific effects can be identified, whereas in the case of a cross-section this is impossible. The panel data models enable us to distinguish between the heterogeneity captured by firm-specific effects and inefficiency. In the following sections the main features of each parametric frontier method are presented.

5.1 Deterministic Frontier Analysis (DFA)

The COLS method is based on a regression analysis. It is a deterministic frontier method since it does not allow for a stochastic error. The cost function is estimated using the OLS technique and then shifted so that all estimated residuals are positive and at least one is zero. The logic of this method was suggested by Winsten (1957). The cost function to be estimated can be written as follows:

$$\begin{aligned} \ln C_i &= c(\mathbf{y}_i, \mathbf{w}_i; \alpha, \boldsymbol{\beta}) + \varepsilon_i \\ &= \alpha + c(\mathbf{y}_i, \mathbf{w}_i; \boldsymbol{\beta}) + \varepsilon_i , \end{aligned} \tag{5.1}$$

where $i = 1,2,\ldots, N$, C represents the observed total cost, $c(\cdot)$ is a suitable functional form,[69] $\mathbf{y}$ the output vector (consisting of output(s) and output characteristics), $\mathbf{w}$ the vector of input prices, and ε the two-sided normally distributed random noise component. Parameter α represents a constant term, while $\boldsymbol{\beta}$ stands for the vector of regression coefficients. The model's parameters, with the exception of the constant term, can be estimated consistently if not efficiently by OLS. After estimating the cost function in Eq.(5.1) with the OLS method, the deterministic cost frontier function by the COLS method is obtained by shifting the cost function in the following way:

$$\ln C_i = \alpha^* + c(\mathbf{y}_i, \mathbf{w}_i; \hat{\boldsymbol{\beta}}) + u_i^*, \tag{5.2}$$

such that $\alpha^* = \hat{\alpha} - |\min\{\hat{\varepsilon}_i\}|$ and $u_i^* = \hat{\varepsilon}_i + |\min\{\hat{\varepsilon}_i\}| \geq 0$ is a once-sided disturbance capturing the effect of cost inefficiency. Proofs of the consistency of the COLS estimator can be found in Greene (1980a). The frontier cost function can thus be viewed as a regression that is in line with the recognition of the theoretical constraint that all observations lie above it. If the distribution of u_i^* were known, the parameters in Eq.(5.1) could be estimated more efficiently by Maximum Likelihood procedure (Greene, 1997).

The cost efficiency score is obtained as:

$$CE_i = \frac{C_i^*}{C_i} = \frac{\exp[\ \alpha^* + c(\mathbf{y}_i, \mathbf{w}_i; \hat{\boldsymbol{\beta}})\ ]}{\exp[\ \alpha^* + c(\mathbf{y}_i, \mathbf{w}_i; \hat{\boldsymbol{\beta}}) + u_i^*\ ]} = \exp(-u_i^*), \tag{5.3}$$

where C_i is the observed total cost and C_i^* is the frontier or minimum cost of the i-th firm. The cost efficiency of the firm is expressed in terms of a score on a scale from 0 to 1 with the frontier firm receiving a score of 1. Alternatively, the cost inefficiency score can be calculated as the reciprocal of the cost efficiency score defined in Eq. (5.3).

Another version of the deterministic frontier was suggested by Afriat (1972) and extended by Richmond (1974). The model is usually referred to as Modified OLS (MOLS). In contrast with the COLS, in the MOLS model the estimated OLS intercept is shifted down by the expectation of inefficiency:

[69] At this point we are not interested in any specific functional form of the cost function. The only requirement is that the cost function is linear in parameters. The choice of the most appropriate functional form will be made in the empirical part of the thesis.

$$\begin{aligned}\ln C_i &= \left(\hat{\alpha} - E[u_i]\right) + c(\mathbf{y}_i, \mathbf{w}_i; \hat{\boldsymbol{\beta}}) + \left(\varepsilon_i + E[u_i]\right) \\ &= \alpha^* + c(\mathbf{y}_i, \mathbf{w}_i; \hat{\boldsymbol{\beta}}) + u_i^*.\end{aligned} \tag{5.4}$$

This, in turn, requires an additional assumption regarding the distribution of the inefficiency component. The MOLS method is a little less orthodox than the COLS method since it is unlikely to result in a full set of negative residuals.

One drawback of the deterministic frontier methods is that they do not allow for stochastic errors and rely heavily on the position of a single most efficient unit. It is assumed that all deviations of observations from the theoretical minimum are attributed solely to the inefficiency of firms. One should note that deviations from the frontier might not be entirely under the control of the firm being studied. However, in the interpretation of the deterministic frontier, for example, an unusually high number of random equipment failures or even bad weather might ultimately appear to the analyst as inefficiency. Moreover, any error or imperfection in the specification of the model or measurement of the variables under consideration could likewise translate into increased inefficiency measures. In contrast with the deterministic frontier, the stochastic frontier specification recognises the possibility of stochastic errors.

5.2 Stochastic Frontier Analysis (SFA)

Stochastic Frontier Analysis allows for inefficiency yet it also acknowledges the fact that random shocks outside the control of producers can affect costs. These models allow for the presence of stochastic errors which embody measurement errors, any other statistical noise, and random variation of the frontier across firms (Greene, 1997). Figure 5.1 provides a basic classification of the stochastic frontier methods. As already noted, we first distinguish between cross-sectional and panel data models. When a cross-sectional model is applied to a panel data set we speak of a pooled model. Panel data models can be further categorised as time-invariant and time-varying inefficiency models. New developments in the field of panel data stochastic frontier models will also be presented.

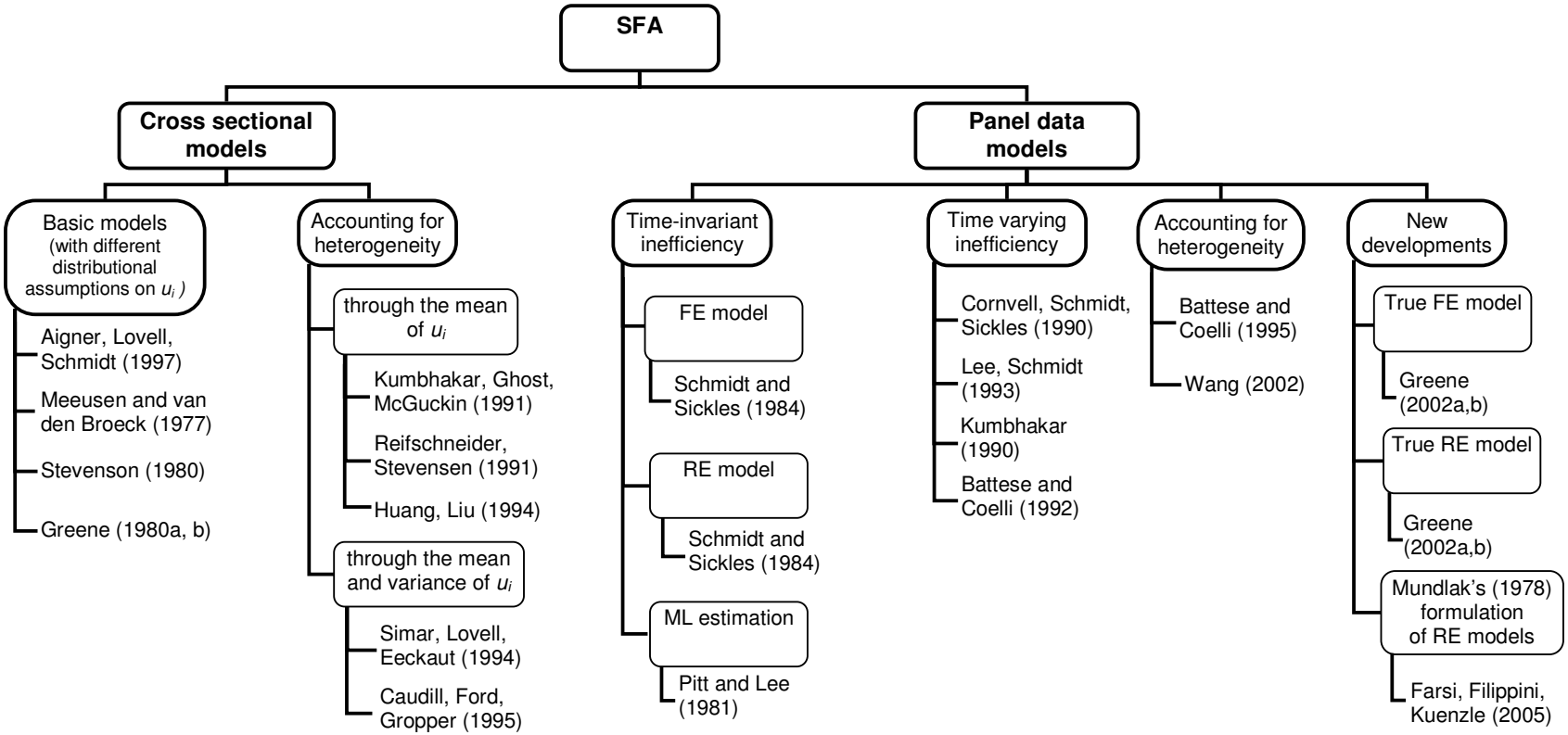

Figure 5.1: Stochastic frontier models

5.2.1 Cross-Sectional Models

5.2.1.1 The Basic Model

Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977) independently introduced stochastic production frontier models. The estimation procedure can be analogously applied to cost frontier models. The stochastic frontier cost function can be written as:

$$\ln C_i = c(\mathbf{y}_i, \mathbf{w}_i; \alpha, \boldsymbol{\beta}) + v_i + u_i \,. \tag{5.5}$$

The error term (ε_i) is now composed of two parts: a stochastic error (v_i), capturing the effect of noise, and a one-sided non-negative disturbance capturing the effect of inefficiency ($u_i \geq 0$). The stochastic frontier cost function is illustrated in Figure 5.2.

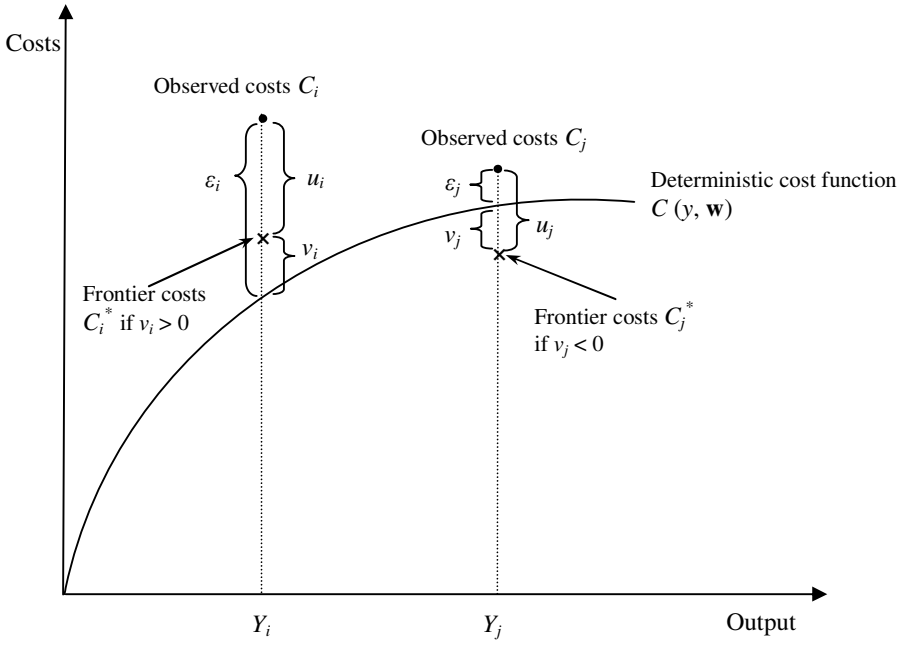

Figure 5.2: Stochastic frontier cost function[70]

[70] A modification of a similar figure in Battese (1992, p.182) in which the stochastic frontier production function is presented.

Two firms, operating at output levels Y_i and Y_j, are considered. They produce given output levels with respective total observed costs C_i and C_j. The deterministic cost frontier is illustrated by the function $C(y, \mathbf{w})$. If we did not account for the stochastic noise and were to perform a deterministic frontier analysis, the inefficiency component would be captured by the terms ε_i and ε_j, respectively. In the stochastic frontier analysis, inefficiency is measured by the terms u_i and u_j while the remaining part of the error term ε is attributed to stochastic errors. For firm i the symmetric error component v_i is positive, implying an unfavourable production environment or the presence of a positive measurement error. On the contrary, firm j operates in a comparatively more favourable environment or there is a negative measurement error present, resulting in a negative value for v_j. Since firms cannot influence the stochastic errors v they are not included in the inefficiency term. Therefore, cost inefficiency is represented by the ratio C_i^*/C_i, where C_i^* represents the frontier of the minimum cost level for the i-th firm.

Accounting for stochastic errors requires the specification of a probability function for the distribution of statistical noise and inefficiencies. To estimate the stochastic cost frontier using the Maximum Likelihood (ML) Method, the following distributional assumptions have to be made:

(i) $v_i \sim \text{iid}\, N(0, \sigma_v^2)$;

(ii) $u_i \sim \text{iid}\, N^+(0, \sigma_u^2)$;[71] and

(iii) v_i and u_i are distributed independently of each other and of the regressors.

Following Kumbhakar and Lovell (2000), this model is referred to as a Normal-Half Normal Model. The individual density functions of v and $u \geq 0$ are respectively:

$$f(v) = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left\{-\frac{v^2}{2\sigma_v^2}\right\}, \tag{5.6}$$

$$f(u) = \frac{2}{\sqrt{2\pi}\sigma_u} \exp\left\{-\frac{u^2}{2\sigma_u^2}\right\}. \tag{5.7}$$

Given the independence assumption, the joint density function of u and v is the product of the two:

[71] Alternative distributional assumptions on u_i will be discussed at the end of this section.

$$f(u,v) = \frac{2}{2\pi\sigma_u\sigma_v}\exp\left\{-\frac{u^2}{2\sigma_u^2}-\frac{v^2}{2\sigma_v^2}\right\}. \tag{5.8}$$

Since $\varepsilon = v+u$, the joint density function for u and ε is:

$$f(u,\varepsilon) = \frac{2}{2\pi\sigma_u\sigma_v}\exp\left\{-\frac{u^2}{2\sigma_u^2}-\frac{(\varepsilon-u)^2}{2\sigma_v^2}\right\}. \tag{5.9}$$

The marginal density function of ε is then:

$$\begin{aligned} f(\varepsilon) &= \int_0^\infty f(u,\varepsilon)du \\ &= \frac{2}{\sqrt{2\pi}\sigma}\cdot\exp\left\{-\frac{\varepsilon^2}{2\sigma^2}\right\}\cdot\left[1-\Phi\left(\frac{-\varepsilon\lambda}{\sigma}\right)\right] \\ &= \frac{2}{\sigma}\cdot\phi\left(\frac{\varepsilon}{\sigma}\right)\cdot\Phi\left(\frac{\varepsilon\lambda}{\sigma}\right), \end{aligned} \tag{5.10}$$

where $\sigma = (\sigma_u^2+\sigma_v^2)^{1/2}$, $\lambda = \sigma_u/\sigma_v$, and $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal cumulative distribution and density functions. The marginal density function $f(\varepsilon)$ is asymmetrically distributed with the mean $E(\varepsilon) = E(u) = \sigma_u\sqrt{2/\pi}$ and the variance $V(\varepsilon) = \sigma_u^2(\pi-2)/\pi+\sigma_v^2$.

Parameter λ provides an indication of the relative contributions of u and v to ε. When λ approaches 0 we are back to the OLS method, whereas when λ goes to infinity we end up with a deterministic frontier with no noise. Coelli (1995) proposed a statistical test of the hypothesis that $\lambda = 0$. The appropriate one-sided likelihood ratio test statistic was shown to be asymptotically distributed as a mixture of χ^2 distributions rather than as a single χ^2 distribution.

This model can be estimated using the Maximum Likelihood estimation method originally proposed by Aigner, Lovell and Schmidt (1977). Using Eq.(5.10), the log likelihood function for a sample of N producers is:

$$\ln L = \text{constant} - N\ln\sigma + \sum_i \ln\Phi\left(\frac{\varepsilon_i\lambda}{\sigma}\right) - \frac{1}{2\sigma^2}\sum_i \varepsilon_i^2 , \tag{5.11}$$

where $\varepsilon_i = \ln C_i - c(\mathbf{y}_i, \mathbf{w}_i; \alpha, \boldsymbol{\beta})$. The log likelihood function can be maximised with respect to the parameters to obtain maximum likelihood estimates of all the parameters (α, $\boldsymbol{\beta}$, σ, λ). However, the focus of the SFA is not on estimating the frontier cost function but rather on the error term, especially the inefficiency component. Therefore, the next step is to obtain estimates of the cost efficiency of each firm. We have estimates of $\varepsilon_i = u_i + v_i$, which obviously contain information on u_i. We can extract or obtain this information from the conditional distribution of u_i given ε_i, which contains whatever information ε_i encompasses concerning u_i. The conditional distribution of u given ε as proposed by Jondrow et al. (1982) is:

$$\begin{aligned} f(u|\varepsilon) &= \frac{f(u,\varepsilon)}{f(\varepsilon)} \\ &= \frac{1}{\sqrt{2\pi}\sigma_*}\exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2}\right\} \Big/ \left[1-\Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right], \end{aligned} \tag{5.12}$$

where $\mu_* = \varepsilon\sigma_u^2/\sigma^2$ and $\sigma_*^2 = \sigma_u^2\sigma_v^2/\sigma^2$. The mean of this distribution can then serve as the point estimator for cost inefficiency u_i:

$$\begin{aligned} E(u_i|\varepsilon_i) &= \mu_{*i} + \sigma_*\left[\frac{\phi(-\mu_{*i}/\sigma_*)}{1-\Phi(-\mu_{*i}/\sigma_*)}\right] \\ &= \sigma_*\left[\frac{\phi(\varepsilon_i\lambda/\sigma)}{1-\Phi(-\varepsilon_i\lambda/\sigma)} + \left(\frac{\varepsilon_i\lambda}{\sigma}\right)\right]. \end{aligned} \tag{5.13}$$

Cost efficiency scores (CE_i) are then calculated using Eq.(5.3). Undesirably, the estimates of cost inefficiency are inconsistent since the variation associated with the distribution of ($u_i|\varepsilon_i$) is independent of i and so the variance of the conditional mean of ($u_i|\varepsilon_i$) for each individual producer does not go to zero as the size of the sample increases (Kumbhakar and Lovell, 2000). However, this is the best that can be achieved with cross-sectional data.

Our analysis so far has been based on the assumption that inefficiency term u_i is distributed as a non-negative half-normal. This distributional assumption is plausible and tractable and

so it is typically employed in empirical work. Other distributional assumptions on the one-side error component u_i can also be used. Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977) suggested the exponential distribution for u_i. Stevenson (1980) introduced a truncated normal distribution of u_i, which is a generalisation of the half-normal distribution. The mean of the underlying normal distribution of inefficiency is allowed to be nonzero, hence $u_i \sim \text{iid}\, N^+(\mu, \sigma_u^2)$. This has the effect of allowing the efficiency distribution to shift to the right, which will allow the mode to move to the right of zero and allow more observations to be further from zero. Moreover, Greene (1980a, 1980b) and Stevenson (1980) assumed that u_i follows a gamma distribution which is a generalisation of the exponential distribution. The latter formulation was further extended by Greene (1990).

We do not intend to go into any more detail here regarding the different assumptions on u_i, but instead address another issue. We focus our attention on different ways to accommodate environmental and non-discretionary factors in the cost efficiency analysis. These factors are considered as exogenous since they are beyond managerial control, but because they can affect the performance of the firm they should be included in the efficiency analysis. We consider different cross-sectional models which can be easily extended to the panel data context.

5.2.1.2 Accounting for Exogenous Factors

For regulated public utilities operating within a network industry, such as water distribution utilities, the size of the service area, population density, type of customers, water treatment, water losses and quality of water are supposedly exogenous factors. These factors are normally included in the model to control for cost differences that occur due to the heterogeneity of output. Let $\mathbf{z}$ denote $L\times 1$ vector of exogenous variables. If $\mathbf{z}$ influences the production process directly, as is the case with network industries, it is appropriate to include $\mathbf{z}$ as a vector of explanatory variables in a stochastic cost frontier (along with $\mathbf{y}$ and $\mathbf{w}$). The model can be rewritten as:

$$\ln C_i = c(\mathbf{y}_i, \mathbf{w}_i, \mathbf{z}_i; \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}) + v_i + u_i, \tag{5.14}$$

where $\boldsymbol{\gamma}$ is a vector of regression coefficients associated with exogenous variables in $\mathbf{z}$. In this case the analysis carried out in the previous section is essentially unchanged. The estimation with the ML procedure requires that the elements of $\mathbf{z}_i$ are uncorrelated with disturbance components v_i and u_i. According to this characterisation, the exogenous variables influence

performance by influencing the production process and not by influencing efficiency. In this way, the frontier cost function is more accurately specified. Variation in efficiency on the other hand is left unexplained by this formulation.

An alternative approach tries to explain variation in efficiency with variation in exogenous variables. In this formulation, environmental and non-discretionary variables influence costs indirectly through its effect on estimated efficiency. Initially, a two-stage procedure was proposed to estimate the model. In the first stage, the inefficiency effects are assumed to be independently and identically distributed in order to use the same approach as in Section 5.2.1.1. The stochastic cost frontier is estimated excluding exogenous variables. It is assumed that elements of $\mathbf{y}_i$ and $\mathbf{w}_i$ are uncorrelated with $\mathbf{z}_i$ since otherwise we get biased estimators due to the omission of $\mathbf{z}_i$. In the second stage, the estimated inefficiencies are regressed against the exogenous variables. In contrast with the previous approach, it is desired for the elements of $\mathbf{z}_i$ to be correlated with u_i. Since now u_i are a function of the exogenous variables, this implies they are not identically distributed as was assumed in the first stage. The two-stage formulation thus has serious econometric flaws.[72]

To overcome this problem, Kumbhakar, Ghost and McGuckin (1991) proposed a single-stage stochastic production frontier model.[73] Likewise, the stochastic cost frontier model can be specified as:

$$\ln C_i = c(\mathbf{y}_i, \mathbf{w}_i; \alpha, \boldsymbol{\beta}) + v_i + u_i, \tag{5.15}$$

$$u_i = \boldsymbol{\gamma}'\mathbf{z}_i + e_i, \tag{5.16}$$

where the cost inefficiency term u_i has a systematic component $\boldsymbol{\gamma}'\mathbf{z}_i$ associated with exogenous variables and a random component e_i. Cost inefficiency is assumed to follow a truncated normal distribution: $u_i \sim N^+(\boldsymbol{\gamma}'\mathbf{z}_i, \sigma_u^2)$. Inserting the expression for u_i in Eq.(5.16) in the cost frontier function in Eq.(5.15) yields:

$$\ln C_i = c(\mathbf{y}_i, \mathbf{w}_i; \alpha, \boldsymbol{\beta}) + v_i + \boldsymbol{\gamma}'\mathbf{z}_i + e_i. \tag{5.17}$$

[72] Wang and Schmidt (2002) argue that if there are any interesting effects to be observed in the second step, then it follows from considerations of omitted variables that the first-step estimators are biased and inconsistent.

[73] The deterministic version of the model was earlier proposed by Deprins and Simar (1989a, b).

This model can be estimated in a single stage by the ML procedure. The log-likelihood function is a straightforward generalisation of that of the truncated normal model introduced by Stevenson (1980) with $\mu_i = \boldsymbol{\gamma}'\mathbf{z}_i$ replacing the constant mean μ. Reifschneider and Stevenson (1991) and Huang and Liu (1994) proposed models similar to those of Kumbhakar, Ghost and McGuckin (1991).[74] Reifschneider and Stevenson (1991) specified the inefficiency term as $u_i = g(\mathbf{z}_i;\boldsymbol{\gamma}) + e_i$, while Huang and Liu (1994) expanded the function $g(\cdot)$ to allow for interactions between elements of $\mathbf{z}_i$ and explanatory variables included in the stochastic frontier function. Battese and Coelli (1995) extended these approaches to accommodate panel data, where $u_{it} = \boldsymbol{\gamma}'\mathbf{z}_{it} + e_{it}$ and $u_{it} \sim N^+(\boldsymbol{\gamma}'\mathbf{z}_{it}, \sigma_u^2)$.

So far, we have relaxed the constant-mean property of the truncated normal distribution and allowed the mean to be a function of exogenous variables. This, in turn, allows inefficiency to depend on exogenous variables. In addition, it is also possible to relax the constant-variance property of the truncated normal distribution (or some alternative distribution) by allowing the variance to be a function of the exogenous variables. Since inefficiency also depends on the variance, this also allows inefficiency to depend on exogenous variables. The latter was done in a model developed independently by Simar, Lovell and Eeckaut (1994) and Caudill, Ford and Gropper (1995). Let us assume that the inefficiency term is specified as:

$$u_i = \exp(\boldsymbol{\gamma}'\mathbf{z}_i) \cdot e_i, \tag{5.18}$$

with e_i being iid with $e_i \geq 0$, $E(e_i) = 1$ and $V(e_i) = \sigma_e^2$. Under these assumptions $u_i \geq 0$ with $E(u_i) = \exp(\boldsymbol{\gamma}'\mathbf{z}_i)$ and $V(u_i) = \exp(2\boldsymbol{\gamma}'\mathbf{z}_i) \cdot \sigma_e^2$. Thus, the variance of u_i is producer-specific. By inserting Eq.(5.18) in Eq.(5.15) the following stochastic cost frontier model is formulated:

$$\begin{aligned} \ln C_i &= c(\mathbf{y}_i, \mathbf{w}_i; \alpha, \boldsymbol{\beta}) + \exp(\boldsymbol{\gamma}'\mathbf{z}_i) \cdot e_i + v_i \\ &= c(\mathbf{y}_i, \mathbf{w}_i; \alpha, \boldsymbol{\beta}) + \exp(\boldsymbol{\gamma}'\mathbf{z}_i) + \varepsilon_i, \end{aligned} \tag{5.19}$$

where $\varepsilon_i = \ln C_i - E(\ln C_i) = v_i + \exp(\boldsymbol{\gamma}'\mathbf{z}_i) \cdot (e_i - 1)$. The ε_is are independently but not identically distributed. If the distribution for e_i is specified, estimators can be obtained with

[74] The models slightly differ in the distributional assumptions and conditions imposed to ensure that $u_i \geq 0$. All three models are estimated by the ML procedure.

the ML techniques. Caudill, Ford and Gropper (1995) derived the log likelihood function for the case in which e_i is exponentially distributed, while Simar, Lovell and Eeckaut (1994) derived the log likelihood function for the gamma and truncated normal distribution (as well as for their exponential and half-normal special cases).

Wang (2002) extended the model of Caudill, Ford and Gropper (1995) combined with the model of Battese and Coelli (1995) to the panel data case. The inefficiency component is assumed to follow a truncated normal distribution, i.e. $u_{it} \sim N^{+}(\mu_{it}, \sigma_{it}^2)$, with $\mu_{it} = \boldsymbol{\gamma}' \mathbf{z}_{it}$ and $\sigma_{it}^2 = \exp(\boldsymbol{\gamma}' \mathbf{z}_{it})$.

The final point to be made is that exogenous variables may belong in the cost frontier or they may belong in the one-sided error component (mean and/or variance of u_i). In most cases, however, it is not obvious whether a certain exogenous variable is a characteristic of production technology or a determinant of inefficiency and a decision has to be made based on the researcher's judgment. This issue was first recognised by Deprins and Simar (1989b).

In the following section we turn to the panel data models which offer much broader possibilities for analysing a firm's inefficiency compared to cross-sectional data and enable us to distinguish between firm-specific effects and inefficiency.

5.2.2 Panel Data Models

In the panel data case, Eq.(5.5) can be rewritten as follows:

$$\ln C_{it} = c(\mathbf{y}_{it}, \mathbf{w}_{it}; \alpha, \boldsymbol{\beta}) + v_{it} + u_{it}, \tag{5.20}$$

where $i = 1,\ldots, N$; $t = 1,\ldots, T$. The cross-sectional model presented in Section 5.2.1.1 can also be employed in the case of a panel data set. We simply consider repeated observations of a given firm as independent observations and thus employ the pooled model. However, in this way we do not exploit the panel aspect of the data.

A panel data set consists of repeated observations on each producer and thus contains more information than a cross-section. Consequently, access to panel data will either result in estimates of cost efficiency with more desirable statistical properties or enable some of the strong distributional assumptions used with cross-sectional data to be relaxed. The Maximum Likelihood estimation requires an assumption that the inefficiency error

component be independent of the regressors, although it might be the case that cost inefficiency is correlated with the input vectors producers choose. Having access to panel data enables us to adapt conventional panel data estimation techniques to the cost inefficiency measurement problem and not all of these techniques rest on strong distributional assumptions.

5.2.2.1 Time-Invariant Cost Inefficiency Models

We first consider panel data cost frontier models in which cost inefficiency is allowed to vary across producers but is assumed to be constant over time. We can formulate this by simply replacing u_{it} in Eq.(5.20) by u_i. In this framework several conventional panel data models can be adapted. The resulting fixed and random-effects models were proposed by Schmidt and Sickles (1984). Another alternative is to extend the Maximum Likelihood model employed in the cross-sectional case to a panel data setup as was done by Pitt and Lee (1981).

5.2.2.1.1 Fixed-Effects Model

If we do not make any distributional assumptions on time-invariant cost inefficiency u_i and also relax the assumption that u_i are uncorrelated with v_{it} and with the regressors, we can use the fixed-effects approach to estimate the following model:

$$\ln C_{it} = \alpha_i + c(\mathbf{y}_{it}, \mathbf{w}_{it}; \boldsymbol{\beta}) + v_{it}, \tag{5.21}$$

where $\alpha_i = \alpha + u_i$ are firm-specific intercepts, $u_i \geq 0$ is time-invariant cost inefficiency and v_{it} represents random statistical noise. We assume that v_{it} are $\text{iid}(0, \sigma_v^2)$ and are uncorrelated with the regressors. The model can be estimated by Least Square Dummy Variables (LSDV) or a 'Within Groups' estimator. The estimate of $\boldsymbol{\beta}$ is obtained by a least squares regression of group mean deviations ($y_{it} - \bar{y}_i$) on explanatory variables. The individual specific dummy variable coefficients can be estimated using the group specific averages of residuals. After estimation the cost frontier intercept is estimated as $\hat{\alpha} = \min_i \{\hat{\alpha}_i\}$ and u_i are estimated from $\hat{u}_i = \hat{\alpha}_i - \hat{\alpha}$, which ensures that all $\hat{u}_i \geq 0$. Efficiency estimation in this model is only with

respect to the 'best' firm in the sample. Efficiency scores in all panel data models are also obtained by Eq.(5.3).

The fixed-effects model has the virtue of simplicity and has nice consistency properties. Under the assumptions of the linear regression model it follows that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$ when $N \to \infty$. The main advantage of the fixed-effects specification is that this estimator is consistent even if the individual effects u_i are correlated with explanatory variables and the random error v_{it}. This is of a vast importance in the network industries where heterogeneity of the output is typically higher than in other industries. Besides the output distributed, one also has to take into account the area size, customer density and several other network characteristics. It may often be the case that some unobserved heterogeneity is present in the specified model and that individual effects may be correlated with explanatory variables. It is thus important to have an estimator that produces unbiased results of coefficients even in the presence of such 'irregularities'. Furthermore, the consistency property does not require the assumption that v_{it} be normally distributed.

However, the fixed-effects model has quite a few shortcomings. The individual effects α_i are each estimated with only T group specific observations and, since T might be small, the estimator of $\boldsymbol{\alpha}$ is inconsistent. This is referred to as the 'incidental parameter' problem in estimating firm-specific effects (Hsiao, 2003). As a result, firm-specific inefficiency estimates are inconsistent. Nonetheless, this inconsistency is not transmitted to $\boldsymbol{\beta}$ since it is not a function of α_i. Also, in contrast to the ML cross-sectional model, the fixed-effects panel data model does provide consistent estimates of firm-specific cost inefficiency when $T \to \infty$ (Kumbhakar and Lovell, 2000).

Further, by the fixed-effects model parameters of the model are estimated from the deviation of the variables from their respective firm-specific means. Therefore, estimation of this model requires that the variables for a given company show enough variation over time. If the within variation is relatively small, the accuracy of the within estimator is limited (Cameron and Trivedi, 2005).

The serious shortcoming of the fixed-effects model is that time-invariant firm characteristics cannot be included in the model as explanatory variables. This implies that the cost inefficiency estimates also capture the effects of all phenomena that vary across producers but are time-invariant for each producer. In network industries many time-invariant characteristics can be found which considerably influence the cost. In the case of water distribution, for example, one could typically mention the area size, morphology of the area,

water resource used, and treatment level. Also, any time-invariant unobserved heterogeneity is captured in the cost inefficiency. This is the reason why in the network industries inefficiency scores estimated by the fixed-effects model are generally found to be considerably higher compared to the other models. This confounding of variation in cost inefficiency with variation in other effects as well as unobserved heterogeneity motivates interest in other panel data models.

5.2.2.1.2 Random-Effects Model

If we do not make any distributional assumptions on u_i but maintain the assumption that u_i are uncorrelated with the v_{it} and with the regressors, we can use the random-effects approach to estimate the following model:

$$\ln C_{it} = \alpha^* + c(\mathbf{y}_{it}, \mathbf{w}_{it}; \boldsymbol{\beta}) + v_{it} + u_i^*, \tag{5.22}$$

where $\alpha^* = \alpha + E(u_i)$ and $E(u_i^*) = E\left[u_i - E(u_i)\right] = 0$. Now u_i are no longer fixed as in the fixed-effects model but are randomly distributed with a constant mean and variance. We do not make any distributional assumptions on the u_i. The model can be estimated with the feasible Generalised Least Square (GLS) method as proposed by Schmidt and Sickles (1984). After the estimation of Eq.(5.22), an estimate of cost inefficiency $\hat{u}_i$ is obtained from the regression residuals as $\hat{u}_i = \hat{u}_i^* - \min_i\{\hat{u}_i^*\} \geq 0$, where:

$$\hat{u}_i^* = (1/T)\sum_t \left[\ln C_{it} - \hat{\alpha}^* - c(\mathbf{y}_{it}, \mathbf{w}_{it}; \hat{\boldsymbol{\beta}})\right]. \tag{5.23}$$

The random-effects panel data model also produces consistent estimates of cost inefficiency when $N \to \infty$ and $T \to \infty$ (Kumbhakar and Lovell, 2000). As opposed to the LSDV estimator, the GLS estimator allows for the inclusion of time-invariant regressors. Consequently, the primary advantage of the random-effects model is that inefficiency estimates do not contain time-invariant firm characteristics. They may, however, still capture the effect of time-invariant, firm-specific unobserved heterogeneity.

The GLS estimator is based on the assumption that firm-specific effects are uncorrelated with the regressors, whereas the LSDV estimator is not. If this assumption does not hold, the GLS estimator may produce biased estimates of the parameters. As already pointed out, this assumption may not hold for network industries since output heterogeneity may result in the

presence of unobserved heterogeneity in the specified model. Nevertheless, the independence assumption can be tested by performing a Hausman-Taylor (1981) test. Moreover, Mundlak (1978) suggested how to overcome this problem and proposed a method to relax the independence assumption between the individual effects and regressors.[75]

5.2.2.1.3 Maximum Likelihood Estimation

The fixed and random-effects models allow us to avoid the strong distributional assumptions made in the cross-sectional frontier models. Nevertheless, if such assumptions are plausible in a panel data context then a Maximum Likelihood estimation is feasible. Pitt and Lee (1981) extended the model of Aigner et al. (1977) to a panel data setup, where the inefficiency term is treated as time-invariant. The resulting Random Effects model is estimated using the Maximum Likelihood estimation method.

The ML estimation of a stochastic cost frontier panel data model in Eq.(5.20) with time-invariant cost inefficiency u_i is, technically speaking, similar to the procedure applied to the cross-sectional data. It is based on the same normal-half normal distributional assumptions. The density function of $u \geq 0$, which is independent of time, is given by Eq.(5.7). The density function of $\mathbf{v} = (v_1, \ldots, v_T)'$, which is time dependent, is given by the following generalisation of Eq.(5.6):

$$f(\mathbf{v}) = \frac{1}{(2\pi)^{T/2}\sigma_v^T} \exp\left\{-\frac{\mathbf{v}'\mathbf{v}}{2\sigma_v^2}\right\}. \tag{5.24}$$

The joint density function of u and $\mathbf{v}$ is used to construct the joint density function of u and $\boldsymbol{\varepsilon} = (v_1 + u, \ldots, v_T + u)'$, from which the log likelihood function for N producers each observed for T time periods is derived. The log likelihood function is then maximised to obtain ML estimates of α, $\boldsymbol{\beta}$, σ_v^2, and σ_u^2. Estimates of producer-specific time-invariant cost inefficiency are obtained as follows:[76]

$$E(u_i|\boldsymbol{\varepsilon}_i) = \mu_{*i} + \sigma_* \left[\frac{\phi(-\mu_{*i}/\sigma_*)}{1-\Phi(-\mu_{*i}/\sigma_*)}\right], \tag{5.25}$$

[75] See Section 5.2.2.3.3.

[76] For details, see Kumbhakar and Lovell (2000).

where $\mu_{*i} = T\sigma_u^2\bar{\varepsilon}_i/(\sigma_v^2 + T\sigma_u^2)$, $\sigma_*^2 = \sigma_u^2\sigma_v^2/(\sigma_v^2 + T\sigma_u^2)$, and $\bar{\varepsilon}_i = \frac{1}{T}\sum_t \varepsilon_{it}$.

The appeal of the ML method is that it should produce more efficient parameter estimates than either LSDV or GLS, but it requires strong distributional assumptions. This technique is widely used in empirical analysis. As in the cross-sectional case, other distributional assumptions on the one-side error component u_i can be utilised as an alternative to the half-normal distribution.

5.2.2.2 Models with Time-Varying Cost Inefficiency

The assumption of time-invariant cost inefficiency is a strong one, particularly in long panels. It might be plausible in a non-competitive operating environment, for example in the case of public utilities that operate in a given service area as local monopolies. However, where competitive pressures are present or we have many time periods it is desirable to relax this assumption. This can be done at the cost of additional parameters to be estimated. Hence, we turn to the panel data cost frontier models where cost efficiency is allowed to vary across producers and over time. These models were first proposed by Cornvell, Schmidt and Sickles (1990) and Kumbhakar (1990).

Cornvell, Schmidt and Sickles (1990) specified the following model:

$$\begin{aligned} \ln C_{it} &= \alpha_t + c(\mathbf{y}_{it}, \mathbf{w}_{it}; \boldsymbol{\beta}) + v_{it} + u_{it} \\ &= \alpha_{it} + c(\mathbf{y}_{it}, \mathbf{w}_{it}; \boldsymbol{\beta}) + v_{it}\,, \end{aligned} \tag{5.26}$$

where α_t is the cost frontier intercept common to all producers in period t and $\alpha_{it} = \alpha + u_{it}$ is the intercept for producer i in period t. Besides obtaining estimates of parameters in $\boldsymbol{\beta}$ and σ_v^2, this requires the estimation of additional $N\times T$ intercepts which is with $N\times T$ observations clearly impossible. This problem was addressed by specifying

$$\alpha_{it} = \theta_{1i} + \theta_{2i}t + \theta_{3i}t^2\,, \tag{5.27}$$

which reduces the number of intercept terms to $3N$. Nevertheless, there is still a lot of parameters to be estimated. The authors propose a fixed-effects and random-effects approach to estimate the model. The fixed-effects model in Eq.(5.26) is estimated by first ignoring u_{it}.

In the second step, the residuals constructed as $\hat{\varepsilon}_{it} = \ln C_{it} - c(\mathbf{y}_{it}, \mathbf{w}_{it}; \hat{\boldsymbol{\beta}})$ are regressed on a constant, t and t^2 for each producer separately to obtain estimates of θ_{1i}, θ_{2i} and θ_{3i}. If N/T is relatively small, another way to estimate α_{it} is to keep u_{it} in Eq.(5.26), estimate θ_{1i} by including producer dummies, and to estimate θ_{2i} and θ_{3i} as coefficients of producer dummies interacted with t and t^2, respectively. After creating estimates of α_{it}, the cost frontier intercept is estimated as $\hat{\alpha}_t = \min_t \{\hat{\alpha}_{it}\}$ and cost inefficiency terms u_{it} are estimated from $\hat{u}_{it} = \hat{\alpha}_{it} - \hat{\alpha}_t$. By construction, in each period at least one producer is found to be fully cost efficient. Cornvell, Schmidt and Sickles (1990) also developed a GLS estimator for a random-effects model. The estimation of intercepts and inefficiencies of this model proceeds as in the fixed-effects approach. The only difference here is that different residuals are used. The advantage of the random-effects model compared to the fixed-effects model is that it can incorporate time-invariant regressors. GLS is also more efficient than the fixed-effects estimator. However, GLS is an inconsistent estimator if cost inefficiencies are correlated with the regressors. For such cases, the authors developed an efficient instrumental variables estimator. Kumbhakar (1990) also developed a maximum likelihood estimator for the model in Eq.(5.26).

Lee and Schmidt (1993) proposed an alternative formulation in which the inefficiency effects for each firm in a different time period is defined as a product of individual inefficiency and time effects:

$$u_{it} = \delta(t) u_i , \tag{5.28}$$

where $\delta(t) = \sum_t \delta_t D_t$ and D_t is a dummy variable for period t. One of the coefficients δ_t is normalised at 1. The cost inefficiency u_i can be estimated either by a fixed- or random-effects model.[77]

Kumbhakar (1990) proposed another model in which u_{it} in Eq.(5.26) is specified as the following function of time:

$$u_{it} = \left[1 + \exp(\gamma t + \delta t^2)\right]^{-1} u_i . \tag{5.29}$$

[77] A generalised method of moments was proposed in Ahn, Lee and Schmidt (1994).

This model requires only two additional parameters to be estimated, γ and δ. The function in parentheses is bounded between 0 and 1 and can be monotonically decreasing or increasing, concave or convex, depending on the values of γ and δ. The model is estimated by the Maximum Likelihood procedure where, apart from the time-varying assumption and two additional parameters to be estimated, other assumptions remain the same as in the Pitt and Lee (1981) time-invariant model.

Battese and Coelli (1992) suggested an alternative to Kumbhakar (1990) in which inefficiency component u_{it} is assumed to be an exponential function of time:

$$u_{it} = \exp\left[(-\eta(t-T)\right]u_i . \tag{5.30}$$

Here only one additional parameter η needs to be estimated. The function in parentheses is positive and decreases (increases) at an increasing rate if $\eta > 0$ ($\eta < 0$) or remains constant if $\eta = 0$. The authors assumed a normal distribution for v_{it} and a truncated normal for u_i and estimated the model using the Maximum Likelihood method. According to Eq.(5.30) inefficiency effects of different firms in any given time period are equal to the same exponential function of the corresponding firm-specific inefficiency effects in the last period of the panel. This implies that the ordering of the firms with respect to the efficiency scores is the same in all time periods which is quite a limiting feature of the model. There is no reason for all the firm-specific deviations to obey the same trajectory. This systematic movement of inefficiency retains a rigid model structure. The model does not account for those situations in which some firms may be initially relatively inefficient but become relatively more efficient in subsequent periods. The Cornvell, Schmidt and Sickles (1990) model does accommodate this possibility and is considered to be more flexible than the Kumbhakar (1990) and Battese and Coelli (1992) specifications. However, this comes at the expense of having many more parameters to estimate.

Besides the time-varying cost efficiency, in long panels it is also desirable to allow for technical change. A time indicator can be included among explanatory variables in a time-varying cost efficiency model enabling one to disentangle the effect of technical change from that of efficiency change (Kumbhakar and Lovell, 2000). The same can be done in the time-invariant cost efficiency models.

As can be seen, several alternatives were proposed to model cost inefficiency. Nevertheless, some issues still need to be properly addressed. One of them is dealing with heterogeneity, which is typically present in the network industries. Excluded variables may result in biased coefficient estimates and inefficiency estimates. By accounting for exogenous factors the

problem is only partially solved since one can control for observed heterogeneity, whereas the problem of unobserved heterogeneity remains. As a result, the unobserved heterogeneity may also produce biased results. Therefore, we now turn to models that have been recently proposed that try to deal with this issue.

5.2.2.3 Recently Proposed Models

The conventional panel data stochastic frontier methods assume that inefficiency is time-invariant. In a lengthy panel, this is likely to be a particularly strong assumption. Moreover, the conventional fixed and random-effects estimators force any time-invariant cross unit heterogeneity into the same term that is being used to capture the inefficiency. It is thus argued that these models fail to distinguish between individual heterogeneity and inefficiency and thus mistakenly measure that heterogeneity as inefficiency. Time-varying inefficiency panel data models relax the unrealistic assumption of unchanging cost inefficiency but do not satisfactory solve the problem of separating heterogeneity and inefficiency. The same holds for those models that account for exogenous factors as they can only control for observed heterogeneity. Nevertheless, not all relevant data are always available and some factors may even be too complex to be properly measurable. This results in unobserved heterogeneity which is beyond the firms' control but may affect their costs significantly.

To deal with the unobserved heterogeneity, the alternative 'true' fixed-effects and 'true' random-effects models proposed by Greene (2002a, 2002b) are considered. Further, Mundlak's (1978) specification of a random-effects model in the stochastic frontier framework as proposed by Farsi, Fillipini and Kuenzle (2005) is considered to avoid possible problems due to the correlation between firm-specific effects and explanatory variables.

5.2.2.3.1 The 'True' Fixed-Effects Model

The motivation for a true fixed-effects model is to treat fixed effects and inefficiency separately and in this way to try to remedy some shortcomings of the conventional fixed-effects model. The latter model can be viewed as the reinterpretation of the linear regression model, while the former is more explicitly built on the stochastic frontier model and uses results that specifically employ the nonlinear specification. There are two issues to be considered, that is the practical problem of computing the fixed-effects estimator and the

bias and inconsistency of the fixed-effects estimator due to the incidental parameters problem.

The true fixed-effects model is specified as follows (Greene, 2002a):

$$\ln C_{it} = \alpha_i + c(\mathbf{y}_{it}, \mathbf{w}_{it}; \boldsymbol{\beta}) + v_{it} + u_{it}. \tag{5.31}$$

Eq.(5.31) can be estimated from the stochastic frontier model using the ML procedure by simply creating dummy variables for each firm. Since the number of dummy variables can be quite large, the method is also known as a 'brute force' approach. However, this approach has seldom been used in the literature. It will be discussed in what follows, after we mention other possible approaches.

The fixed-effects model discussed in Section 5.2.2.1.1 is based on a linear regression. There, by using group mean deviations, fixed effects are removed from the model. Consequently, the slope estimator $\boldsymbol{\beta}$ is not a function of fixed effects and is thus consistent (unlike the estimator of fixed effects $\boldsymbol{\alpha}$). The same holds for nonlinear models in which there are minimal sufficient statistics for the individual effects $\boldsymbol{\alpha}$. In these cases, the log likelihood conditioned on the sufficient statistics is a function of $\boldsymbol{\beta}$ that is free of fixed effects. However, this cannot be done for the true fixed-effects model so the method is not useful in our case.

Heckman and MaCurdy (1980) suggested a 'zig-zag' approach to maximisation of the log likelihood function, dummy variable coefficients and all. In the first step, a known set of fixed-effects coefficients $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_N)'$ is assumed. From the log likelihood function conditioned on these values $(\log L \mid a_1,\ldots,a_N)$, the estimation of $\boldsymbol{\beta}$ is straightforward. In the second step, with a given estimate of $\boldsymbol{\beta}$ obtained in the first step (denoted by $\mathbf{b}$), the conditional log likelihood function for each α_i $(\log L_i \mid \mathbf{b})$ is now a known function. Maximising this function is straightforward but somewhat tedious since it has to be done for each i. This two-step optimisation presumes iterating back and forth between these two estimators until convergence is achieved. In principle, this approach could be adapted with any model. However, there is no guarantee that this back and forth procedure will converge to the true maximum of the log likelihood function because the Hessian is not block diagonal. Whether either estimator is consistent in dimension N, even if T is large, depends on the initial estimator being consistent (Greene, 2002a).

Polachek and Yoon (1994, 1996) essentially applied this approach to a fixed-effects stochastic frontier model, which was the first study to fully implement true fixed effects in the stochastic frontier setting. They specified the frontier and constructed a likelihood function based on an exponential distribution rather than the half-normal. In the first step they proposed to estimate a fixed-effects panel data model with the Least Square Dummy Variables as in Section 5.2.2.1.1., then the computation of the fixed effects by the within group residuals. In the second step, true fixed effects α_i in the log likelihood are replaced by these estimates a_i and the resulting function is maximised with respect to the small number of remaining parameters ($\boldsymbol{\beta}$, σ, λ). Then, fixed effects are recomputed by the same method and returned to the log likelihood function to re-estimate the other parameters. These steps are repeated until convergence is reached. While the initial OLS estimator of $\boldsymbol{\beta}$ is consistent, the subsequent estimators, which are functions of the estimated fixed effects a_i, are not because of the incidental parameters problem. More will be said on this below. The initial OLS estimator obeys the familiar results for the linear regression model, but the second step MLE does not since the likelihood function is not the sum of squares. Also, the asymptotic standard errors of the estimators are found to be underestimated. The differences between the OLS and ML estimators are, however, extremely small. Since the authors were interested in the structural parameters of the model they did not carry on analysing technical inefficiency terms.

An alternative approach to computing the fixed-effects estimator is by the direct maximisation proposed by Greene (2001). The maximisation of an unconditional log likelihood function is done by 'brute force'. This approach is feasible even in the presence of possibly thousands of nuisance parameters. A nonlinear model is defined by the density of ε_{it}, which for the fixed-effects stochastic cost frontier model is defined as:

$$f(\varepsilon_{it}) = \frac{2}{\sigma} \cdot \phi\left(\frac{\varepsilon_{it}}{\sigma}\right) \cdot \Phi\left(\frac{\varepsilon_{it}\lambda}{\sigma}\right), \tag{5.32}$$

where $\varepsilon_{it} = \ln C_{it} - \alpha_i - c(\mathbf{y}_{it}, \mathbf{w}_{it}; \boldsymbol{\beta}) = u_{it} + v_{it}$. A set of group dummy variables α_i is created and included in the model. It is assumed that v_{it} is normally distributed and u_{it} follows a half-normal distribution. The unconditional likelihood function, L, is constructed as follows:

$$L = \prod_{i=1}^{N} \prod_{t=1}^{T} f(\varepsilon_{it}) . \tag{5.33}$$

Then, the log likelihood function, log$L$ is maximised by 'brute force'. The gradient of the log likelihood $\mathbf{g}$ and the Hessian $\mathbf{H}$ are provided in Greene (2002a). Newton's method is used to produce estimates of parameters in each iteration in the following way:

$$\begin{pmatrix}\hat{\boldsymbol{\gamma}}\\ \hat{\boldsymbol{\alpha}}\end{pmatrix}_k = \begin{pmatrix}\hat{\boldsymbol{\gamma}}\\ \hat{\boldsymbol{\alpha}}\end{pmatrix}_{k-1} - \mathbf{H}_{k-1}^{-1}\,\mathbf{g}_{k-1} = \begin{pmatrix}\hat{\boldsymbol{\gamma}}\\ \hat{\boldsymbol{\alpha}}\end{pmatrix}_{k-1} + \begin{pmatrix}\boldsymbol{\Delta}_{\gamma}\\ \boldsymbol{\Delta}_{\alpha}\end{pmatrix}_{k-1}, \tag{5.34}$$

where subscript k indicates the updated value, subscript k-1 indicates a computation at the current value and $\boldsymbol{\Delta}$ stands for the update. The full Hessian $\mathbf{H}$ is of the dimension $(K_{\gamma}+N)\times(K_{\gamma}+N)$, where K_{γ} stands for the number of parameters in $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \sigma, \lambda)$. However, the difficulty of this approach is not in the computation, as may appear. Greene (2001) demonstrated that neither update vector requires the storage or inversion of a $(K_{\gamma}+N)\times(K_{\gamma}+N)$ matrix; each is a function of the sums of scalars and $K_{\gamma}\times 1$ vectors of first derivatives and mixed second-order derivatives.[78] The practical implication is that calculation of the fixed-effects model is a computation only of order K_{γ}. Storage requirements for $\boldsymbol{\alpha}$ and $\boldsymbol{\Delta}$ are linear in N, not quadratic, which is well within the capacity of current computers. Hence, with this approach it is possible to directly compute both the joint maximisers of the log likelihood and the appropriate submatrix of the inverse of the analytic second-order derivatives for estimating asymptotic standard errors.

What remains to be addressed is a statistical problem related to the fixed-effects estimator, namely the incidental parameter problem. The incidental parameter problem is a persistent bias that arises in nonlinear fixed-effects models when the number of periods is small. With a small T many fixed-effects estimators of model parameters are inconsistent and subject to a small sample bias.[79] Beyond the theoretical and methodological results there is almost no systematic analysis of this problem. Greene (2002a, 2005) uses a Monte Carlo analysis to provide some empirical econometric evidence of the severity of the incidental parameter problem in the case of an estimator for the stochastic frontier model.[80] He also analyses how systematic biases in the parameter estimates (if they exist) are transmitted to estimates of the inefficiency scores, which are of primary interest in the stochastic frontier analysis.

---

[78] Similar holds for the asymptotic variances and covariances.

[79] The inconsistency results from the fact that the asymptotic variance of the Maximum Likelihood estimator does not converge to zero as N increases.

[80] The data were taken from the Commercial Bank Holding Company Database maintained by the Chicago Federal Reserve Bank. A random sample of 500 banks from a total of over 5000 US commercial banks was used to estimate the Cobb-Douglas cost frontier function over the 1996-2000 period.

Greene (2002a) discovers that for the structural coefficients in the model the biases in the slope estimators are quite moderate, especially in comparison to some other models like probit, logit or ordered probit. The economies of scale bias is estimated with a very small bias (0.48%), which is smaller than the estimated sampling variation of the estimator itself (roughly 7%). In contrast, the estimator of the constant term seems to be widely underestimated (in some cases the bias is found to be -300% or more). Overall, the deviations of the regression parameters (with the exception of the constant term biases) are found to be small, in particular given the small T (5 years). The bias appears to be toward zero, not away from it as in some other models. Greene (2005) finds that the force of the incidental parameters problem actually shows up in the variance estimators, not in the slope estimators. The estimate of σ appears to have absorbed the force of inconsistency and is considerably overestimated. A similar result appears for λ, but towards rather than away from zero, indicating that the true fixed-effects model does not perform so well after all. Since σ and λ are crucial parameters in the computation of inefficiency estimates, this leads us to expect some large biases in these estimators. The overestimation error of inefficiency estimates is found to be about 25%. Greene (2005) also compares these results with the conventional fixed-effects model. He discovers similar descriptive statistics for the two sets of estimates. However, the correlation of the inefficiency estimates obtained by the true fixed-effects and regression-based fixed-effects model is very weak. Nonetheless, Greene (2002b) concludes that the pessimism about the fixed-effects estimator is overstated.

5.2.2.3.2 The 'True' Random-Effects Model

The random-effects model introduced in Section 5.2.2.1.2 parallels the linear regression model. It has already been discussed that this model is unable to properly distinguish between time-invariant heterogeneity and inefficiency since both effects are captured by the same term. Therefore, (Greene, 2002b) specifies the true random-effects model in the following way:

$$\ln C_{it} = \alpha + c(\mathbf{y}_{it}, \mathbf{w}_{it}; \boldsymbol{\beta}) + \omega_i + v_{it} + u_{it}, \tag{5.35}$$

where ω_i is a time-invariant, firm-specific random effect meant to capture cross-firm heterogeneity. The difference between this formulation and the true fixed-effects model is the additional assumption that ω_i and all other terms in the model are uncorrelated. The model seems to have a three-part disturbance which raises questions of identification. This

interpretation would be misleading as the model actually has a two-part composed error, ω_i and $\varepsilon_{it} = u_{it} + v_{it}$. This is an ordinary random-effects model, albeit one in which the time-varying component ε_{it} has an asymmetric distribution. Conditioned on ω_i, the T observations for firm i are independent so the joint density for the T observations is:

$$f(\varepsilon_{i1},...,\varepsilon_{iT}|\omega_i) = \prod_{t=1}^{T} \frac{2}{\sigma} \cdot \phi\left(\frac{\varepsilon_{it}}{\sigma}\right) \cdot \Phi\left(\frac{\varepsilon_{it}\lambda}{\sigma}\right), \tag{5.36}$$

where $\varepsilon_{it} = \ln C_{it} - (\alpha + \omega_i) - c(\mathbf{y}_{it}, \mathbf{w}_{it}; \boldsymbol{\beta})$. In order to be able to estimate the model parameters, it is necessary to integrate the heterogeneity out of the log likelihood. The unconditional joint density is obtained as:

$$L_i = f(\varepsilon_{i1},...,\varepsilon_{iT}) = \int_{\omega_i} \prod_{t=1}^{T} \frac{2}{\sigma} \phi\left(\frac{\varepsilon_{it}}{\sigma}\right) \Phi\left(\frac{\varepsilon_{it}\lambda}{\sigma}\right) g(\omega_i)\, d\omega_i . \tag{5.37}$$

The log likelihood, $\Sigma_i \log L_i$, can be then maximised with respect to the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma, \lambda)$ and any additional parameters characterising the distribution of ω_i that appear in the maximand. However, the maximisation problem just stated is not solvable because there is no closed form for this integral. By rewriting Eq.(5.37) in the equivalent form:

$$L_i = f(\varepsilon_{i1},...,\varepsilon_{iT}) = E_{\omega_i}\left[\prod_{t=1}^{T} \frac{2}{\sigma} \phi\left(\frac{\varepsilon_{it}}{\sigma}\right) \Phi\left(\frac{\varepsilon_{it}\lambda}{\sigma}\right)\right], \tag{5.38}$$

the log likelihood can be computed by simulation. Averaging Eq.(5.38) over sufficient draws from the distribution of ω_i will produce a sufficiently accurate approximation of the integral to allow an estimation of the parameters. The simulated log likelihood is:

$$\log L_S(\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda, \sigma, \theta) = \sum_{i=1}^{N} \log \frac{1}{R} \sum_{t=1}^{T} \left[\prod_{t=1}^{T} \frac{2}{\sigma} \phi\left(\frac{\varepsilon_{it}|\omega_{ir}}{\sigma}\right) \Phi\left(\frac{\varepsilon_{it}\lambda|\omega_{ir}}{\sigma}\right)\right], \tag{5.39}$$

where θ is used for the parameters in the distribution of ω_i, and ω_{ir} is the r-th simulated draw for observation i. If ω_i is normally distributed, then θ is its standard deviation. Estimation of the true random-effects model can be extremely time consuming. In order to achieve a reasonable approximation to the true likelihood function, a large number of random draws is

required. The process can be accelerated by using draws such as Halton sequences. The use of Halton sequences can reduce the number of draws required by a factor of five or ten.

Greene (2002a, b) compared the results of the true random effects, true fixed effects, Pit and Lee (1981) and cross-sectional models and found that all models produce similar parameter estimates of the stochastic cost frontier, but the correlation between the cost inefficiency estimates is quite loose.[81] He finds that the regression- and likelihood-based treatments of inefficiency bring striking differences in the results. These differences might be undetected if one focuses only on descriptive statistics of the inefficiency estimates. What the nature and source of these differences is remains to be discovered in future research.

Due to the novelty of the true random-effects model only a few applications of this model can be found so far. Farsi, Filippini and Greene (2005) used true random effects to estimate the cost inefficiency of Swiss railway companies, while Farsi, Filippini and Kuenzle (2005, 2006) used this model in the case of Swiss nursing homes and regional bus companies, respectively. In all three cases, the performance of the true random-effects model was compared to the other panel data stochastic frontier models examined in this chapter. Only a moderate correlation between the cost inefficiency estimates resulting from different models was found.

The advantage of the true random-effects over the true fixed-effects model is that time-invariant explanatory variables can be included in the model. The problem of these two models, however, is that any time-invariant or persistent component of inefficiency is completely absorbed in the firm-specific constant term. Whereas the earlier fixed and random-effects models tend to overestimate the inefficiency component, it is possible that the latter two forms will underestimate it. Whether those time-invariant effects belong to unobserved heterogeneity or inefficiency is debatable. Clearly, it is not obvious on inspection how one should treat time-invariant effects in a data set. How this issue is handled has a large influence on the findings. Ultimately, $(\alpha_i + u_{it} + v_{it})$ or alternatively $(\omega_i + u_{it} + v_{it})$ contains both firm-specific heterogeneity and inefficiency and both might have time-invariant and time-varying elements. Unfortunately, there is no perfect way to disentangle them based on observed data.

[81] The data were taken from the Commercial Bank Holding Company Database mentioned earlier (see previous footnote). Greene (2003) also reports similar findings for the WHO panel data set on the country effectiveness of the delivery of health care. Different specifications of the stochastic production frontier model bring substantial changes to the technical efficiency estimates.

5.2.2.3.3 Mundlak's Formulation

When unobserved heterogeneity, captured by the inefficiency term, is correlated with explanatory variables, the random-effects panel data models are affected by 'heterogeneity bias'.[82] In contrast, the fixed-effects estimator produces consistent estimates but has other shortcomings. As noted, time-invariant variables cannot be included in the model so these effects are contained in the firm-specific terms. This in turn results in very high inefficiency estimates. An additional drawback is the 'incidental parameter' problem. Since *T* is usually small, the estimates of individual effects may incur large errors which directly affect inefficiency estimates.

Mundlak (1978) thus proposed an alternative specification of the RE model that controls for the correlation between unobserved heterogeneity and explanatory variables. In this way 'heterogeneity bias' is avoided and inefficiency estimates are adjusted for the heterogeneity that is correlated with explanatory variables. In the stochastic frontier context this approach has only been recently used by Farsi, Fillipini and Kuenzle (2005).

Mundlak (1978) suggested modelling the correlation of firm-specific effects α_i with explanatory variables in an auxiliary regression given by:

$$\alpha_i = \alpha_0 + \boldsymbol{\gamma}'\bar{\mathbf{x}}_i + \delta_i, \tag{5.40}$$

where $\bar{\mathbf{x}}_i = (1/T)\sum_1^T \mathbf{x}_{it}$ and $\delta_i \sim \text{iid}\, N(0, \sigma_\delta^2)$. Here, $\mathbf{x_i}$ stands for the vector of all explanatory variables (output, output characteristics, and input prices)[83] and $\boldsymbol{\gamma}$ is the vector of corresponding coefficients. In this formulation it is assumed that unobserved heterogeneity is correlated with the group-means of the explanatory variables. The individual effects of Eq.(5.31) are divided into two components; the first part can be explained by the exogenous variables, whereas the remaining part δ_i is orthogonal to the explanatory variables. The resulting GLS estimator of the RE model is identical to the 'within' estimator of the FE model, thus unbiased.[84]

[82] The term 'heterogeneity bias' refers to the bias caused by the correlation between individual effects and regressors in the general RE model (Farsi, Fillipini and Kuenzle, 2005).

[83] The variables are considered to be properly transformed (e.g., in logarithms).

[84] For proof of the identity, see Hsiao (2003), pp. 44-46.

Mundlak's (1978) formulation can be applied to the stochastic frontier panel data model by incorporating Eq.(5.40) in Eq.(5.21), which results in:

$$\ln C_{it} = \alpha_0 + \mathbf{x}_{it}'\boldsymbol{\beta} + \boldsymbol{\gamma}'\bar{\mathbf{x}}_i + \delta_i + v_{it} . \tag{5.41}$$

We use the same method described in Section 5.2.2.1.2 to estimate a RE model with δ_i playing the role of u_i. The resulting GLS estimator, which is equivalent to the within estimator, is unbiased. At the same time, we manage to separate (some of the) unobserved heterogeneity which is correlated with explanatory variables from inefficiency. Since we used the RE model, time-invariant factors can also be included in the model. Mundlak's (1978) formulation can also be employed in the other RE models, i.e. in the Maximum Likelihood model by Pitt and Lee (1981) and Greene's (2002a, 2002b) 'true' RE model. These models assume an asymmetric error term and, since they are nonlinear and thus estimated by ML procedure, the equivalence argument with 'within' estimator does not strictly apply. Nevertheless, the 'heterogeneity bias' is expected to be at least partly avoided (Farsi, Fillipini and Kuenzle, 2005).

5.2.3 Summary

We have analysed different parametric methods for estimating cost inefficiency. The main stochastic frontier models and their corresponding features are summarised in Table 5.1. Some models need strong distributional assumptions (e.g. models based on the ML estimation) and/or impose non-realistic assumptions on the inefficiency term (e.g. time-invariant inefficiency). As discussed, cross-sectional models need to be estimated by the ML procedure and do not allow for the firm-specific factors. Therefore, we have put the main focus on the panel data models which offer a much broader set of possibilities for modelling inefficiency. Among other things, we established that some panel data models are unable to distinguish between (unobserved) heterogeneity and inefficiency. For that reason, we introduced Mundlak's (1978) formulation of the random-effects model as proposed by Farsi, Fillipini and Kuenzle (2005) that controls for the correlation between unobserved heterogeneity and explanatory variables. In this way inefficiency estimates are adjusted for the heterogeneity that is correlated with explanatory variables. We also considered recently developed true fixed- and true random-effects models by Greene (2002a, 2002b) which capture the effects of time-invariant unobserved heterogeneity with a separate term. Based on the observed data, one cannot conclude with certainty whether those time-invariant effects belong to unobserved heterogeneity or inefficiency. The choice of appropriate model

is also based on the researcher's belief whether there is some time-invariant unobserved heterogeneity in the model or whether the inefficiency does not in fact vary over time.

So far, we have not put much stress on the issue of consistency of the efficiency estimates resulting from different stochastic frontier models. The results can be sensitive to the model specification, choice of variables and functional form employed. It is thus important to analyse the consistency of the inefficiency scores obtained from different models, especially if SFA is supposed to serve as a benchmarking tool in the price regulation. If different methods produce very different results, we cannot use them directly in economic policy-making. We will turn to this issue and discuss it in more detail in the empirical part of the thesis, more specifically in Section 7.1.

Table 5.1: Summary of the different stochastic frontier models

Authors	Error structure specification	Estimation method	Distribution of u_i and assumptions on u_i
I. Cross sectional models			
Aigner, Lovell and Schmidt (1977)	$\varepsilon_i = v_i + u_i$	ML	$u_i \sim \text{iid}\, N^+(0,\sigma_u^2)$ mutual independence between u_i, v_i and the regressors
Kumbhakar, Ghost and McGuckin (1991)	$u_i = \boldsymbol{\gamma}'\mathbf{z}_i + e_i$	ML	$u_i \sim N^+(\boldsymbol{\gamma}'\mathbf{z_i},\sigma_u^2)$ mutual independence between u_i, v_i and the regressors
II. Panel data models			
Schmidt and Sickles (1984)	$\varepsilon_{it} = v_{it} + u_i$ $\hat{u}_i = \hat{\alpha}_i - \min_i\{\hat{\alpha}_i\}$	FE (LSDV)	time-invariant inefficiency (u_i)
Schmidt and Sickles (1984)	$\varepsilon_{it} = v_{it} + u_i$ $\hat{u}_i = \hat{u}_i^* - \min_i\{\hat{u}_i^*\}$	RE (GLS)	u_i time-invariant, u_i are uncorrelated with v_{it} and the regressors
Pitt and Lee (1981)	$\varepsilon_{it} = v_{it} + u_i$	ML	$u_i \sim \text{iid}\, N^+(0,\sigma_u^2)$ u_{it}, v_{it} and the regressors are mutually independent
Cornvell, Schmidt and Sickles (1990)	$\varepsilon_{it} = v_{it} + u_{it}$ $\alpha_{it} = \theta_{1i} + \theta_{2i}t + \theta_{3i}t^2$ $\hat{u}_{it} = \hat{\alpha}_{it} - \min_i\{\hat{\alpha}_{it}\}$	FE / RE / EIV	no assumptions / u_{it}, v_{it} and the regressors are mutually independent / u_{it} correlated with the regressors
Lee and Schmidt (1993)	$\varepsilon_{it} = v_{it} + u_{it}$ $u_{it} = \delta(t)u_i$, $\delta(t) = \sum_t \delta_t D_t$	FE / RE	no assumptions / mutual independence between u_i, v_i and the regressors
Kumbhakar (1990)	$\varepsilon_{it} = v_{it} + u_{it}$ $u_{it} = \left[1 + \exp(\gamma t + \delta t^2)\right]^{-1} u_i$	ML	$u_i \sim \text{iid}\, N^+(0,\sigma_u^2)$ mutual independence
Battese and Coelli (1992)	$\varepsilon_{it} = v_{it} + u_{it}$ $u_{it} = \exp\left[(-\eta(t-T)\right]u_i$	ML	$u_i \sim N^+(\mu,\sigma_u^2)$ mutual independence
Battese and Coelli (1995)	$\varepsilon_{it} = v_{it} + u_{it}$ $u_{it} = \boldsymbol{\gamma}'\mathbf{z_{it}} + e_{it}$	ML	$u_{it} \sim N^+(\boldsymbol{\gamma}'\mathbf{z_{it}},\sigma_u^2)$ mutual independence
Wang (2002)	$\varepsilon_{it} = v_{it} + u_{it}$ $u_{it} = \boldsymbol{\gamma}'\mathbf{z_{it}} + e_{it}$	ML	$u_{it} \sim N^+(\mu_{it},\sigma_{it}^2)$ $\mu_{it} = \boldsymbol{\gamma}'\mathbf{z_{it}}$, $\sigma_{it}^2 = \exp(\boldsymbol{\gamma}'\mathbf{z_{it}})$ mutual independence
Mundlak's (1978) formulation	$\alpha_i = \alpha_0 + u_i = \alpha_0 + \boldsymbol{\gamma}'\bar{\mathbf{x}}_i + \delta_i$ $\varepsilon_{it} = \delta_i + v_{it}$ $\hat{\delta}_i = \hat{\delta}_i^* - \min_i\{\hat{\delta}_i^*\}$	RE	inefficiency term δ_i is uncorrelated with v_{it} and the regressors
Greene (2002a, b)	$\varepsilon_{it} = v_{it} + u_{it}$, group dummy variables α_i	'Brute force' ML	$u_{it} \sim \text{iid}\, N^+(0,\sigma_u^2)$ mutual independence
Greene (2002a, b)	$e_{it} = \omega_i + v_{it} + u_{it}$ $= \omega_i + \varepsilon_{it}$	Simulated ML	$u_{it} \sim \text{iid}\, N^+(0,\sigma_u^2)$ $\omega_i \sim \text{iid}\, N(0,\sigma_\omega^2)$ mutual independence

6 Estimation of the Cost Frontier Function for Slovenian Water Distribution Utilities

The findings from the economic and econometric theory described in previous chapters are applied to the Slovenian water industry with the goal of estimating the cost frontier function. First, a review of the literature on previous studies of water companies will be provided. Studies covering an estimation of the cost (frontier) function, economies of density, economies of scale and cost efficiency will be considered. The previous empirical findings will then serve as a starting point for specifying the stochastic cost frontier model for Slovenian water distribution companies. Before estimating the model, a panel data sample of water distribution utilities will be described together with the variables included in the model. Since the estimation of the cost frontier function requires a specification of a functional form, a choice of the most appropriate functional form will be made. Several parametric frontier models will be employed when estimating the (stochastic) frontier total cost function. Based on the estimated results, cost efficiency scores and economies of density and scale in Slovenian water distribution utilities will be obtained in the following chapter.

6.1 Literature Review of Cost Studies of Water Distribution Companies

In the literature we find two types of studies on the costs of water distribution companies: (i) studies estimating the cost function and economies of output density, customer density and economies of scale;[85] and (ii) studies estimating the cost frontier function and cost efficiency. In what follows, we provide a short review of the most relevant papers covering sample description, model specification, functional form, variables included in the cost function, the method of estimation and the results. The review is fundamentally given in a chronological order; there is only one exception made in order to discuss studies referring to the same country in the one place. The definitions of economies of scale and density in Section 3.5.2 are based on a cost model that includes the output delivered, number of customers and area size (or alternatively, network length) in specifying the cost function. In this case, one can distinguish between the economies of output density, customer density and scale. However,

[85] More accurately, the reviewed studies estimate economies of size. See Section 3.5.2 for the distinction between economies of scale and size. Nevertheless, no study uses the latter expression so we will also stick to economies of scale.

the empirical studies are found to be quite heterogeneous with respect to how they measure economies of scale and density. Some of them only use output as explanatory variables, some use output and customers or output and area size, while others use all three variables. Consequently, the results obtained regarding economies of scale and density are not fully comparable. Also, in the case of a variable cost function some estimates refer to the short run while others refer to the log run. We will pay close attention to these issues when interpreting the results from different studies. For the literature review, we will keep the definitions proposed by the authors unless they do not correspond to the theoretical findings in Section 3.5.2. It is recalled that, as a general rule, one does not speak of short-run economies of scale.

Kim and Clark (1988) examined the multi-product nature of water supply relative to economies of scale and scope.[86] The data used in the study came from a cross-section of 60 water utilities in the United States for 1973, collected mainly by the US Environmental Protection Agency. The water utility is viewed as a multi-product firm providing residential and non-residential (i.e., industrial, commercial, wholesale and other uses) services. Thus, in estimating the total cost function two outputs are considered: the volumes of water delivered to residential and non-residential users. To take the spatial variation of demand into account service distance is included, which is the distance from the treatment plant to the service area. Due to the capital-intensive nature of the operation, a capacity utilisation rate measured by the load factor of a water system is incorporated in the model. The input prices included in the model are the price of labour, the price of capital and the price of energy.[87] The translog multi-product total cost function is estimated jointly with the cost share equations forming a multivariate regression system, subject to the parameter restrictions imposed by symmetry and linear homogeneity in input prices.[88] No significant economies of scale in the utility's overall operation are discovered, where overall economies of scale for the sample mean are estimated to be 0.99. Small utilities exhibit quite marked economies of scale (1.33),

[86] Kim and Clark (1988) also pointed out that a number of studies have been conducted to estimate the water supply cost function, for example, Crain and Zardkoohi (1978), Bruggink (1982), Feigenbaum and Teeples (1983). Unfortunately, the authors ascertain that most of the previous studies suffer from severe shortcomings in the methodology and specification employed. Also, Teeples and Glyer (1987) compared their model to those in previous studies by Crain and Zardkoohi (1978), Bruggink (1982) and Feigenbaum and Teeples (1983). They argued that the differing conclusions can be put down to the model restrictions implicit in these earlier papers.

[87] The price of capital is the percentage interest rate on long-term debt plus 2%, which is an average depreciation rate.

[88] Zellner's iterative seemingly unrelated regression (SUR) method is used. The choice of the translog is somewhat surprising due to the small sample size.

while large utilities exhibit moderate diseconomies of scale (0.88). The utilities are found on the whole to enjoy considerable economies of scale for non-residential water supply, but suffer from diseconomies in residential supply. The utilities also experience economies of scope associated with the joint production of the two services (0.166). It is estimated that the costs of providing residential and non-residential services separately are about 17% higher than the costs of providing them jointly. The shortcoming of this study is that it does not consider output characteristics in estimating the total cost function. Due to the excluded variables of number of customers and area size, the results may be biased.

Bhattacharyya et al. (1995) used a stochastic frontier cost function to specify the costs and inefficiency of publicly- and privately-owned urban water utilities in terms of their different ownership structures and firm-specific characteristics. The data were obtained from a 1992 survey of water utilities in cities serving a population of over 25,000 in the US. The number of utilities in the sample is 221, of which 190 are publicly-owned and 31 privately-owned firms. A translog functional form is employed to estimate the variable cost function. The explanatory variables used are the output (total quantity of water sales), input prices (price of energy, labour and material), the stock of capital and the network variables. The capital stock is measured as the residual of the revenue less the variable cost divided by the opportunity cost of capital, which can be viewed as a very rough approximation of the actual capital stock. The opportunity cost of capital is defined as the sum of the average depreciation rate and the long-term interest rate. Network configuration variables include the different types of water sources used (a dummy for surface water only and a dummy for a mix of surface and underground water), the total quality of water produced and the total system loss. The error term is composed of a random noise and cost inefficiency term. Both the mean and variance of inefficiency are specified in the model as functions of firm-specific factors. These factors are the ownership dummy variable, the number of emergency breakdowns, the length of distribution pipelines and a dummy variable distinguishing those companies that are only residential suppliers from those that supply both residential and commercial water. The translog variable cost function, the cost share equations and equations specifying the error terms (including the inefficiency term) constitute a system of equations that is estimated simultaneously using the Iterative SUR method. A two-step estimation procedure is conducted. The advantage of the two-step method compared to the single-step ML estimation method is that it does not require strong distributional assumptions regarding the stochastic term. However, the two-step procedure has some serious econometric flaws which results in biased and inconsistent estimators.[89] The results obtained confirm the presence of firm-specific inefficiency effects where both private and public utilities are shown to be

[89] Recall the discussion in Section 5.2.1.2.

significantly cost inefficient. The mean cost inefficiency is estimated to be 11.1%. The public water utilities on average outperform the private water companies; the estimated mean cost inefficiencies are 9.8% and 18.7%, respectively. The results show that when the operation is small privately-owned water utilities are comparatively more efficient. Public water utilities are comparatively more efficient when the scale of operation is large. Returns to density, at the mean, for the privately-owned utilities are estimated to be 1.25 and 0.93 for publicly-owned utilities. While private companies are operating with increasing returns to density and are, given the level of their capital stock, underutilising their capacities, publicly-owned firms experience diminishing returns to density. These results refer to the short run since the numerator of the expression used to calculate economies of (output) density is not adjusted in accordance with Eq.(3.40). Again, the estimated economies of density may be biased since output characteristics are not included in the model. Another important thing to note is that the coefficient of the capital stock has a positive sign. This occurs quite often in the applied literature, although it contradicts the cost theory.[90] One possible reason for this in this particular case may be the poorly specified capital stock variable.

Stewart (1993) estimated the operating expenditure of water supply companies to identify the impact of operating conditions on the costs of companies and any differences in company costs that may not be attributable to variations in operating conditions (i.e., cost inefficiencies). The data collected by the OFWAT provide information for the 32 water companies of England and Wales for 1992/93. The dependent variable modelled is the total water operating expenditure (OPEX) which corresponds to the variable cost.[91] A number of plausible models are examined. A preferred model is derived within which OPEX is found to be related to the volume of water delivered, the length of the distribution system, the amount of pumping required and the proportion of water delivered to measured non-household properties.[92] Very serious weaknesses of this model specification are that it does not include input prices and the capital stock. It is thus questionable whether the function obtained can be actually regarded as the variable cost function. The Cobb-Douglas functional form is employed. Econometric modelling techniques are used with (C)OLS as the main method and

[90] See Section 3.3.2.

[91] Operating expenditures (OPEX) and capital expenditures (CAPEX) sum up to total expenditures (TOTEX).

[92] Due to the very high correlation between the explanatory variables the volume of water delivered, the length of the distribution system, the volume of water put into the distribution system and the number of properties billed (correlation coefficients in excess of 0.97 between any pair) not all of the variables were included in the model. Other variables like the proportion of groundwater, proportion of water from river abstraction, various treatment variables, mains bursts, and ratio of peak to average water delivered were also considered.

SFA as the alternative method, which is viewed as an examination of the robustness of the (C)OLS results. The cost elasticity of the volume of water delivered is estimated to be 0.57 and that of the length of the main to be 0.38. Based on these results, economies of utilisation or short-run economies of output density are estimated to be 1.75 while, as already pointed out, one cannot speak of short-run economies of scale. Considerable variation across companies as regards the extent of cost inefficiency is discovered. The mean sample efficiency is estimated to be 83.6% or, alternatively, the inverse, which is the mean cost inefficiency, equals 19.7%.[93] The estimated cost reductions needed to achieve the best practice vary between 0% for the most efficient company to 33% for the least efficient company. SFA is used to examine the sensitivity of the COLS results.[94] The sample is, however, too small to obtain reliable estimates by employing SFA. The (C)OLS results were employed by the OFWAT in the price-regulation process, i.e. in setting company-specific price caps. This may seem a doubtful decision since, due to the identified weaknesses of the model, i.e. the small sample size used, non-inclusion of input prices and non-inclusion of the capital stock in specifying the variable cost function, the results could be seriously biased.

Ashton (2000) considers water and sewerage companies as integrated firms and investigates water and sewerage costs together. The relative efficiency of 10 privatised regional UK water and sewerage companies between 1987 and 1997 is estimated from the stochastic frontier cost function. The panel data set is unbalanced and contains a total of 92 firm observations. A one-component fixed-effects panel data model is used to estimate a variable cost function from which distribution-free firm-specific estimates of cost efficiency are derived. Thus, the paper uses the fixed-effect stochastic frontier panel data model proposed by Schmidt and Sickles (1984).[95] The translog functional form is used as a representation of productive technology. Again, cost in this study is defined as operating cost only, excluding the effects of depreciation and infrastructural renewal. Explanatory variables are the level of output, price of labour, price of consumables and price of other costs. The level of output is proxied by the overall number of households connected to the distribution system and by applying some additional adjustments. The price of labour is defined as the level of staff cost

[93] The cost efficiency of UK water distribution companies was analysed in a number of papers that followed. Cubbin and Tzanidakis (1998) and Thanassoulis (2000) employed a similar model to Stewart (1993). They compared the COLS results to the results obtained by the DEA method. The same shortcomings as in the model proposed by Stewart (1993) can be identified.

[94] The coefficient estimates of the stochastic cost function and ranking of the companies are almost identical to those obtained by the OLS model, but the estimated inefficiency scores are very different and depend on the distributional assumption used in the SFA (exponential or truncated, while the half-normal model did not converge).

[95] See Section 5.2.2.1.

divided by the number of full-time-equivalent employees. The price of consumables is expressed as the level of spending on consumable inputs (including power, materials and taxes) divided by the level of fixed assets. The price of other costs, incorporating service charges and other direct costs, is outlined as other costs divided by fixed assets. The partial derivative of cost with respect to the output is estimated to be 0.466 which indicates that substantial economies of output density (2.15) are observable in the water and sewerage industry in the short run. Again, it is noted that one does not speak of short-run economies of scale. Overall operating cost efficiency is estimated to be 84%. A moderate level of dispersion in operational cost efficiency is recorded. Since output characteristics are not included in the model, the inefficiency may comprise both diversity of the operating environment as well as differences in performance. Another important variable missing in the estimation of the variable cost function is the capital stock. Nonetheless, due to the use of the fixed-effects model the problem of biased results is probably less serious than in the previous study.

Estache and Rossi (2000) compare the performance of water companies in Asia and the Pacific region. Hence, international frontier benchmarking is conducted. The data published by the Asian Development Bank cover 50 firms surveyed in 1995. Due to some missing values, the final sample consists of 44 companies. The log-log variable cost frontier function is estimated using the COLS and SFA methods. The costs comprise operational and maintenance costs. The explanatory variables included in the model are the price of labour, number of clients, population density in the area served, number of connections, percentage of residential sales in total sales, number of hours of water availability (quality variable) and a dummy variable for concessioned firms. The average efficiency score obtained by the COLS and SFA methods are 0.733 and 0.639, respectively. This is somewhat surprising since it would be expected that the COLS method would produce lower efficiency scores compared to the SFA method, simply due to the fact that the error term in the former case might not only include an inefficiency effect. However, with just 44 observations and possibly some data comparability issues we can hardly claim that the results are reliable, with this holding particularly for the SFA method. The correlation coefficient of 0.861 between the ranks resulting from the COLS and SFA methods indicates there is some consistency between the two approaches. Besides the small sample size used in the study, other shortcomings can be identified. Problems of data comparability are usually more severe in international studies since the heterogeneity of utilities is typically larger. There are also some problems in specification of the variable cost function. The capital stock is left out of the model and output is also not incorporated in the cost function. In the latter case it may be argued that the number of customers can serve as a proxy for the output distributed.

Finally, not all input prices are included in the model (e.g., the price of material and price of energy).

Fabbri and Fraquelli (2000) analyse costs, economies of scale and economies of output density in the Italian water industry. The cross-sectional dataset consists of 150 water companies observed in 1991. These companies represent 3% of the water companies operating in Italy; however, they account for nearly 50% of the volume supplied. The size of the average firm in the sample is thus much bigger than the average size of all Italian water companies. A translog total cost function is used, which is jointly estimated with the cost share equations using Zellner's SUR technique. The authors state they are estimating a hedonic cost function but in fact they use a general specification of the translog cost function.[96] The chosen explanatory variables are the volume of water delivered, the price of labour, the price of energy and the price of capital-material, along with the following output characteristic variables: the number of consumers, a proxy for population density obtained as the ratio between the population served and the length of pipelines, the cost of water input purchased and the treatment costs as a percentage of total costs.[97] The price of capital-material is computed by dividing the sum of depreciation and the cost of material by the length of the network. The estimated cost elasticities at the sample means suggest that economies of output density are present, while economies of scale are not confirmed. Returns to output density are found to equal 1.58 at the expansion point, 14.3 at the minimum point and 0.90 at the maximum. Returns to scale are found to be high at the minimum point (2.38), at the expansion point they are not significantly different from 1 (0.99), while at the maximum point there are instead diseconomies of scale (0.68). Since area size is not included in the model, the estimated economies of scale do not entirely correspond to the definition in Eq.(3.39).

Antonioli and Filippini (2001) also explore economies of scale and density in the Italian water industry. The difference compared to Fabbri and Fraquelli (2000) is that in this paper a panel data set is used and the variable cost function is estimated. The panel consists of 32 water distribution firms over the 1991-1995 period. The total variable cost equals the sum of

[96] Recall the discussion on the hedonic vs. general specification of the cost function in Section 4.1.8. Besides the general specification of the translog function, the authors also estimate the translog cost function without output characteristics and a Cobb-Douglas cost function with and without output characteristics. Since the general translog cost function is proven to be the preferred specification, we only provide results for this function. The variables are normalised around their own sample means. Results for other functional forms are pretty much in line with the general translog specification.

[97] As water purchased and treatment costs may take on zero values, the Box-Cox transformation has been applied.

direct costs and labour costs. The explanatory variables employed are the amount of water distributed, the price of labour, the number of customers, the length of the pipes (as a proxy for area size), the percentage of water losses, the number of water wells (as a proxy for the capital stock), a treatment dummy variable and a time variable to capture shifts in technology. To avoid the multicollinearity problem a Cobb-Douglas functional form is employed. The variable cost function is estimated using the OLS and random-effects model. Due to the time-invariance of some explanatory variables, the fixed-effects model was disregarded. The estimated functions are well behaved with most of the parameter estimates carrying the expected sign and being statistically significant. The paper obtains a negative sign for the capital stock coefficient which is in line with theoretical expectations. However, the coefficient is not significant. The inclusion of the output, number of customers and the size of the service area in the cost function allows for the distinguishing of economies of output density, economies of customer density and economies of scale. The results obtained refer to the long run since the numerator of the expression to calculate respective economies is adjusted according to Eq.(3.40). Since returns to scale are estimated to be 0.95, the results based on the random-effects model suggest the presence of weak diseconomies of scale. On the contrary, there are economies of output and customer density, with the estimates being equal to 1.46 and 1.16, respectively. Since the random-effects model is unable to control for unobserved heterogeneity the results may be biased.

Mizutani and Urakami (2001) analyse the optimal size and economies of scale and output density for Japanese water supply companies. The observations used in the study comprise 112 water supply organisations for the 1994 fiscal year. As far as the functional form is concerned, the log-log, the translog and the translog with a hedonic specification of the output are used. The cost measure used is the total cost which is estimated as the function of output (total volume of water delivered), factor prices (labour, energy, material and capital prices) and network characteristics (network length and utilisation rate).[98] In addition, in the hedonic specification of the cost function the output measure is specified as a function of the output and output quality measures (treatment or purifier level, household ratio, non-dam water acquisition index, and non-underground water index).[99] To estimate the cost models with input share equations the SUR method is used. The preferable model is found to be the

[98] The material price is obtained by dividing repair expenditures by fixed assets. The capital price is defined as the sum of depreciation rate and the interest rate on short-term bonds held by governments. The utilisation rate is defined as the ratio of the daily water delivery volume to the designated volume of water-intake. Since not all variables were available they had to be estimated. Such variables are the network length and the energy consumption needed for the energy price calculation.

[99] The treatment level is defined by dividing the clear water volume by a designated volume of water-intake.

translog with a hedonic specification of the cost function. The authors confirm economies of output density. On the contrary, no evidence of scale economies is found. It should be noted that the definition of economies of scale in this study does not fully correspond to Eq.(3.39) since the number of customers is not included in the model. For average-sized water supply companies, returns to output density based on the estimated hedonic specification of the cost function are 1.10, while returns to scale are reported to be 0.92.[100] The same functional form is used to estimate the optimal size of a company, i.e. the size with the minimum average cost (in terms of output and network length). The minimum average cost is found to be at the point of 261,084,000 m^3 of water supplied and 1,221 km of network length.

Garcia and Thomas (2001) examine the cost structure of French municipal water utilities. The sample is composed of 55 water utilities from the Bordeaux region for the years 1995 to 1997. The Generalised Method of Moments (GMM) procedure is used to estimate the system of variable cost and input cost shares.[101] A multi-product translog cost function is employed. Several measures of returns are assessed as well as the economies of scope associated with the joint production of water delivered to final customers and water losses. Variable cost is defined as the sum of total operating and maintenance cost. The following explanatory variables are used: the output variables (volume sold to final customers and water losses), factor prices (labour price, energy price, material price) and technical variables. Water losses are measured as the difference between the water distributed and the water sold. The price of material is obtained by dividing material costs with the volume of water distributed, whereby the material costs include heterogeneous categories of costs such as stocking, maintenance work and subcontracting. The technical variables used are the number of customers, the number of municipalities supplied (as a proxy for area size), and several variables representing the existing capital stock: network length, production capacity, stocking and the pumping capacity. Estimated economies of scope at the variables sample mean are positive (0.237), indicating that there are potential gains in the production of water losses (undesirable output) jointly with the water sold to final customers (desirable output). A possible explanation of this is that the costs associated with network repairs and maintenance in order to decrease water losses are higher than the costs involved in satisfying customer demand by simply increasing water production. Further, returns estimates at the sample mean show that, in the short run, there are economies of output density (1.14) as well as

[100] Companies are grouped in five categories according to their size (measured by the amount of water supplied). For all groups and all functional forms applied similar results are obtained, i.e. economies of output density and diseconomies of scale are found.

[101] In the linear regression model, the GMM method is equivalent to the Instrumental Variables (IV) method.

economies of customer density (1.05). In the long run, economies of output density are found (1.21), while there are no longer any economies of customer density present (0.87). Finally, (long-run) scale economies are estimated to be 1.002.

Table 6.1 summarises all of the abovementioned studies – the data set used, method and functional form employed, and reports the main findings regarding the economies of density, economies of scale and cost (in)efficiencies of the water distribution companies. Based on the studies reviewed, several shortcomings can be identified:[102]

- a small sample size is used;
- the non-inclusion of input prices (or non-inclusion of all input prices) in the cost function;
- the non-inclusion of output characteristics in the cost function;
- studies estimating variable cost should incorporate capital stock in the model;
- not distinguishing in a precise way between the short-run and long-run economies of scale and density; and
- several deficiencies in the methods chosen to estimate cost inefficiency can be found.

The consequences of the weaknesses identified above may be quite severe. One of the main issues to be stressed is the non-inclusion of output characteristics to account for output heterogeneity and the non-inclusion of some other relevant variables in the model. Due to the excluded variables, the estimated coefficients of the cost function could be seriously biased. In particular, the exclusion of output characteristics which are typically highly correlated with the output results in biased estimates of economies of scale and density. Further, cost inefficiency estimates may be sensitive to the stochastic frontier method employed. Another main problem that failed to be recognised by the authors of the reviewed studies is that none of the stochastic frontier methods employed is able to differentiate between unobserved heterogeneity and inefficiency. Consequently, unobserved heterogeneity is simply attributed to cost inefficiency. This is particularly undesirable in the case of network industries where heterogeneity is usually found to be rather high. There can be found several environmental and non-discretionary factors that are beyond managerial control, but can affect the performance of the firms. Some of these factors can be observed and included in the cost function to control for the cost differences that occur merely due to differences in operating environment. However, the problem of unobserved heterogeneity may still be present. It is thus essential to be able to make a distinction between different sources of cost differentials, namely between unobserved heterogeneity and inefficiency. This may, in turn, have very important implications from the regulatory point of view. If differences in costs are

[102] Of course, not all of the studies suffer from (all of) the above listed shortcomings.

attributed to differences in the cost inefficiency, the policy implications will be completely different than if these differences are attributed to differences in operating environment.

With respect to the previous studies, the contribution of this study is that it notably improves on stochastic frontier methods used to estimate cost inefficiency. It explicitly recognises the problem of unobserved heterogeneity in measuring the cost inefficiency and tries to address this problem in several ways. The most recent findings on stochastic frontier estimation methods are considered. The novelty is to use the TFE model as proposed by Greene (2002a) and the RE model with Mundlak's formulation as recently proposed by Farsi, Filippini and Kuenzle (2005). Additionally, we are interested in analysing to what extent the estimated inefficiency scores are found to be sensitive to the SFA method employed. Thus, the consistency of the results obtained by employing different stochastic frontier methods will as well be analysed. We will be especially interested to find out whether the established inconsistency can be explained by different ability of the models to distinguish unobserved heterogeneity from inefficiency.

It should be also noted that this is one of the first studies to measure inefficiency in the network industries from Central and Eastern European transition countries and the first such study relating to the water sector in these countries.[103] Only recently have the transition countries recognised the importance of this issue. Thus, for economic policy decision-making it is important to have at least some indication of the presence of inefficiencies in these countries' network industries. The results can also provide a useful input when designing the appropriate regulatory framework.

[103] By becoming a new EU member state on May 1 2004, the transition process of Slovenian economy has been formally completed. Thus, we can now speak of ex-transition countries.

Table 6.1: Summary of the findings from the literature review

Author(s) of the paper	Data sample	Model and functional form	Method of estimation or calculation	Estimated economies of scale (size)	Estimated economies of density	Estimated cost efficiency
Kim and Clark (1988)	60 US water utilities in 1973	Translog multi-product TC function	SUR method	0.992 (sample average)[1]	/	/
Bhattacharyya et al. (1995)	221 US water utilities from a 1992 survey	Translog VC function	SFA (SUR and two-step estimation)	/	1.246 (E_{OD}, SR, private); 0.932 (E_{OD}, SR, public)[2]; group means	0.901 (average; public more efficient)
Stewart (1993)	32 UK water companies in 1992/93	Log-log VC function[3]	OLS and COLS method	/	1.75 (E_{OD}, SR)	0.836
Ashton (2000)	10 UK water and sewerage companies between 1987-1997	Translog VC function[3]	SFA (fixed-effects panel data model)	/	2.15 (E_{OD}, SR, sample average)	0.84
Estache and Rossi (2000)	44 water utilities in Asia observed in 1995	Log-log VC function	COLS SFA (pooled model)	/	/	0.733 (COLS) 0.639 (SFA)
Fabbri and Fraquelli (2000)	150 Italian water utilities observed in 1991	Translog TC function	SUR method	0.99 (sample average)	1.58 (E_{OD}, sample average)	/
Antonioli and Filippini (2001)	32 Italian water utilities between 1991-1995	Log-log VC function	OLS and RE panel data model	0.95 (LR)	1.46 (E_{OD}, LR) 1.16 (E_{CD}, LR)	/
Mizutani and Urakami (2001)	112 Japanese water companies in 1994	Log-log, translog, hedonic TC function	SUR method	0.921 (sample average)	1.103 (E_{OD}, sample average)	/
Garcia and Thomas (2001)	55 French water utilities between 1995-1997	Multi-product translog VC function	GMM (IV method), SUR method	1.002 (LR, sample average)	1.21 (E_{OD}, LR); 0.87 (E_{CD}, LR);[4] sample average	/

[1] Economies of scope in providing residential and non-residential services jointly are also confirmed ($E_{SCOPE} = 0.166$).
[2] E_{OD} stands for economies of output density, while E_{CD} stands for economies of customer density. SR stands for short run, while LR denotes long run.
[3] Instead of VC, the authors actually speak of operating expenditures (OPEX).
[4] In the short run, $E_{OD} = 1.14$ and $E_{CD} = 1.05$. Economies of scope associated with the joint production of water delivered to final customers and water losses are also found ($E_{SCOPE} = 0.24$).

6.2 Model Specification and Methodology

The main purpose of water distribution companies is to produce water of sufficient quality from a resource (groundwater or surface water) that may necessitate preliminary treatments to make drinking water safe, and to distribute water by continuously adapting supply to daily demand while preserving water quality during its transportation in the transmission pipelines and distribution mains. Underlying technological constraints clearly play an important role in constructing the water cost function. Municipal water supply covers all operations from resource extraction through to consumer taps (Garcia and Thomas, 2001).

The water-production process consists of the following activities (Fabbri and Fraquelli, 2000, and Garcia and Thomas, 2001):

- *Production and treatment* covers the operations of water extraction (from groundwater or surface water) through to preliminary treatment in treatment plants (disinfection, iron removal, filtering, softening);
- *Transfer* is the carrying of water from production facilities to transmission pipelines that can be gravity or pump-operated if a ground storage system is employed;
- *Storage* of water in facilities such as water tanks and water towers;
- *Pressurisation* of water pipelines, either by a gravity main system or with the help of a pumping station; and
- *Distribution* of water to final customers through distribution mains and customer service lines; it also includes quality monitoring and metering.

A special feature of water utilities worth mentioning is that the water delivered to final customers is obtained from raw, untreated water, which has no acquisition costs. This is why it is not treated in the same way as other inputs (labour, energy, material and capital). The production and delivery processes of water distribution companies are highly dependent on their capital stock consisting of pumps, network pipelines, storage facilities and other facilities. The technical environment in which water utilities operate can be very different. In the production activity, water utilities can be distinguished based on the water source used: groundwater or surface water. Groundwater implies higher drilling and pumping costs, whereas treatment costs are usually higher with surface water. Differences in average costs can also be found in the distribution process, depending on the size of the service area, population density, customer mix (share of household vs. non-household customers), water leakages from the network system etc. Therefore, in order to deal with such heterogeneity it is necessary to incorporate in the cost function, along with the factor prices and the output

(i.e., water delivered to final customers), variables that represent output characteristics and differences in the environmental conditions of the water utilities.

The costs of operating a water distribution system are the costs of building and maintaining the water system (wells and springs, pumps, treatment facilities, distribution pipes and other facilities), and of measuring and billing water. The main factors influencing the cost of water distribution companies are:

a) the total water sold;
b) the input prices;
c) the total number of customers served;
d) the type of consumers;
e) the customer density;
f) the size and morphology of the distribution area;
g) the length of distribution pipes;
h) the water resource (underground water or surface water);
i) water losses from the distribution system;
j) the load factor (ratio between the average and maximum demand of water); and
k) the water treatment needed.

For a specification of the cost model, we consider a water distribution company which uses three inputs, labour, capital and material, to distribute a single output to a number of customers within its service area. The number of customers and the network size can be considered as output characteristic variables. The output characteristics are included as explanatory variables to control for the cost differences that occur merely due to the heterogeneity of output.

In estimating a cost function, a decision has to be made whether to employ a short-run or a long-run cost function. From the literature review on water distribution companies it can be seen that some studies estimated a variable cost function, while others decided for the total cost function (see Table 6.1). The decision is generally based on our belief as to whether the utilities use all inputs at their optimal levels. If it is assumed that utilities are in a long-run static equilibrium with respect to all inputs employed and that they minimise total cost, a total cost function is utilised. On the other hand, if it is believed that utilities do not use a certain input at its optimal level, a variable cost function is utilised. In the latter case, capital is typically considered to be a fixed input since time is required to adjust it to its optimal level through the investment process. Besides that, the decision on which concept to employ may be influenced by the availability of the data and econometric considerations.

It is the case that with water distribution utilities the capital is long-lived and adjustments to a change in water demand are costly. However, over the last few decades we have not witnessed any notable changes in water demand in Slovenia so the demand for water can be considered to have been relatively stable. There is thus no reason to believe that utilities are considerably deviating from the optimal level of capital employed. Therefore, the adoption of a long-run concept seems an appropriate choice. Another reason we prefer the total cost function is to avoid possible manipulations related to the allocation of costs to different cost categories. It may happen that some companies may view a certain cost item as a variable cost, while others consider it to be a capital cost item (e.g., maintenance as opposed to investment expenditure). Also, if only variable costs were considered in the efficiency analysis we would implicitly allow for some inefficiency with respect to the capital costs. If utilities realised that capital costs are exempt from assessing inefficiency they might try to move some cost items from variable to capital costs in order to be perceived as more efficient.

Hence, based on the above discussion a decision is made to employ a total cost function, which can be written as:

$$C = C(Q, P_L, P_M, P_K, CU, AS, D_{LOSL}, D_{TREAT}, D_S, D_U, T), \tag{6.1}$$

where C represents total cost and Q is the output represented by the total cubic metres of water delivered. P_L, P_M, and P_K are the price of labour, the price of material and the price of capital, respectively. CU stands for the number of customers served, while AS is the size of the service area. D_{LOSL} is a dummy variable of water losses bearing a value of 1 if the firm has low percentage of water losses, and a 0 value otherwise. D_{TREAT} is a dummy variable for water treatment and takes on a value of 1 if the firm distributes water that has to be treated chemically before distribution and a 0 value otherwise. Treatment is necessary in a situation where, from a medical point of view, the quality of the water does not reach a predefined standard and it is therefore unsuitable for drinking. Water distribution utilities can use different water resources: surface water, underground water or a mix of both sources. D_S represents a dummy variable for surface water only and D_U is a dummy variable for underground water only. Finally, T is a time variable, which captures the shift in technology.

Estimation of the cost function requires a specification of a functional form. In Chapter 4 we considered several functional forms and discussed the criteria for choosing a functional form prior to estimation. To review, a functional form is considered to be appropriate if it satisfies properties indicated by economic theory, because of its flexibility, possibility and ease of statistical estimation and application, and consistency with empirical facts. Following these

criteria we selected three functional forms: Cobb-Douglas, general translog and hedonic translog function. This is also in line with the studies examined in the literature review (see Table 6.1).

Other functional forms were not found to be appropriate for estimating the cost function with output characteristics since they did not satisfy one or more criteria for choosing a functional form. To meet theoretical consistency conditions, the cost function has to be nondecreasing in output, linearly homogeneous in input prices and nondecreasing and concave in input prices. Therefore, the quadratic cost function is disregarded since it is not linearly homogeneous in input prices nor can homogeneity be imposed without sacrificing the flexibility of the form. The same holds for the quadratic mean of order one. Further, the generalised Leontief functional form is not locally flexible whereas the locally flexible version of a generalised Leontief is defined only for a single output firm and does not allow the inclusion of output characteristics. The hybrid Diewert multi-product cost function is also not a locally flexible functional form and, from the point of view of the present study, the fact that it imposes constant returns to scale is particularly unwelcome. Finally, a general Box-Cox model and the models nested within it are eliminated from the set of possible alternatives since they do not meet the criterion regarding the ease of statistical estimation. Since the general Box-Cox model is nonlinear in parameters it has to be estimated by a nonlinear maximum likelihood technique. In a preliminary analysis, in fact, an attempt was made to estimate the general Box-Cox model as well as some of its nested versions. However, the Nonlinear Least Squares (NLS) method used to estimate these models did not converge so these models had to be eliminated.

Hence, we end up with three possible functional forms all nested within the translog functional form, namely the Cobb-Douglas, the general translog and the hedonic translog cost function. The Cobb-Douglas is not a locally flexible form but, because of its simplicity of application and the clearness of interpretation of its parameters, it is widely used and will therefore be tested against the translog. The major limit of the Cobb-Douglas functional form is that the estimated values of the economies of scale and density do not vary with the size of the firms in the sample but are assumed to be constant. Generally speaking, the translog cost function, which is a more flexible functional form, appears to be an appropriate functional form for answering questions about economies of scale and density. In the general translog specification, output characteristics are included directly in the cost function, i.e. in a general manner without imposing any restrictions. An alternative way to include output characteristics is by estimating the translog with a hedonic specification of output. Its appeal is that far fewer parameters need to be estimated compared to the general translog specification. However, the hedonic translog cost function also has some shortcomings. It

requires the separability of the arguments in each hedonic aggregator from all other arguments in the cost function. As a result of this restrictive assumption, the translog cost function with a hedonic output specification is no longer capable of providing a second-order approximation to any unknown arbitrary separable cost function.

By applying the translog functional form, Eq.(6.1) can be rewritten as:

$$\begin{aligned}
\ln\frac{C_{it}}{PK_{it}} &= \ln a + b_Q \ln Q_{it} + b_{CU} \ln CU_{it} + b_{AS} \ln AS_{it} + c_{PL} \ln\frac{PL_{it}}{PK_{it}} + c_{PM} \ln\frac{PM_{it}}{PK_{it}} \\
&+ \frac{1}{2} b_{Q,Q} \ln Q_{it} \ln Q_{it} + \frac{1}{2} b_{CU,CU} \ln CU_{it} \ln CU_{it} + \frac{1}{2} b_{AS,AS} \ln AS_{it} \ln AS_{it} \\
&+ b_{Q,CU} \ln Q_{it} \ln CU_{it} + b_{Q,AS} \ln Q_{it} \ln AS_{it} + b_{CU,AS} \ln CU_{it} \ln AS_{it} \\
&+ \frac{1}{2} c_{PL,PL} \ln\frac{PL_{it}}{PK_{it}} \ln\frac{PL_{it}}{PK_{it}} + \frac{1}{2} c_{PM,PM} \ln\frac{PM_{it}}{PK_{it}} \ln\frac{PM_{it}}{PK_{it}} + c_{PL,PM} \ln\frac{PL_{it}}{PK_{it}} \ln\frac{PM_{it}}{PK_{it}} \\
&+ d_{PL,Q} \ln\frac{PL_{it}}{PK_{it}} \ln Q_{it} + d_{PL,CU} \ln\frac{PL_{it}}{PK_{it}} \ln CU_{it} + d_{PL,AS} \ln\frac{PL_{it}}{PK_{it}} \ln AS_{it} \\
&+ d_{PM,Q} \ln\frac{PM_{it}}{PK_{it}} \ln Q_{it} + d_{PM,CU} \ln\frac{PM_{it}}{PK_{it}} \ln CU_{it} + d_{PM,AS} \ln\frac{PM_{it}}{PK_{it}} \ln AS_{it} \\
&+ g_{LOSL} D_{LOSL} + g_{TREAT} D_{TREAT} + g_S D_S + g_U D_U + h_T T + \varepsilon_{it}, \qquad (6.2)
\end{aligned}$$

with ε_{it} the error term. Notice that the normalisation of cost and input prices by one of the input prices is used to impose linear homogeneity in input prices. Hence, the total cost, the price of labour and the price of material are divided by the price of capital. In what follows, we use the following notation for the normalised variables: C^*, PL^* and PM^*. Other properties of the cost function remain to be verified after the translog cost function is estimated.

The Cobb-Douglas functional form is estimated by imposing the following restrictions on the translog cost function in Eq.(6.2):

$$\begin{aligned}
&b_{Q,Q} = b_{CU,CU} = b_{AS,AS} = b_{Q,CU} = b_{Q,AS} = b_{CU,AS} = 0, \\
&c_{PL,PL} = c_{PM,PM} = c_{PL,PM} = 0, \qquad (6.3) \\
&d_{PL,Q} = d_{PL,CU} = d_{PL,AS} = d_{PM,Q} = d_{PM,CU} = d_{PM,AS} = 0.
\end{aligned}$$

In the translog specification in Eq.(6.2), output characteristics are included directly in the cost function, i.e. in a general manner without imposing any restrictions. An alternative way to include output characteristics is by estimating the translog with a hedonic specification of

output. Consistent with Section 4.1.8, the following Cobb-Douglas output aggregator function is employed:

$$\ln Y(Q,CU,AS) = h_Q \ln Q_{it} + h_{CU} \ln CU_{it} + h_{AS} \ln AS_{it}, \tag{6.4}$$

where $h_Q = 1$. In Eq.(6.4) only two output characteristics are taken into account, namely the number of customers served and the size of the service area. Unfortunately, the other output attributes appearing in the cost function specified by Eq.(6.1) are all dummy variables and could not be incorporated into the output aggregator function but were included directly in the cost function. The translog cost function in Eq.(6.2) now takes the following form:

$$\ln C^* = C(Y, P_L^*, P_M^*, D_{LOSL}, D_{TREAT}, D_S, D_U, T). \tag{6.5}$$

By substituting Eq.(6.4) in Eq.(6.5) a general translog function is obtained. A hedonic translog cost function can then be estimated through a set of nonlinear constraints on the parameters of the general translog function. In our case, the following restrictions are imposed on Eq.(6.5):[104]

$$h_{AS} = \frac{b_{AS}}{b_Q} = \frac{b_{AS,AS}}{b_{Q,AS}} = \frac{b_{Q,AS}}{b_{Q,Q}} = \frac{b_{CU,AS}}{b_{Q,CU}} = \frac{d_{PL,AS}}{d_{PL,Q}} = \frac{d_{PM,AS}}{d_{PM,Q}}, \tag{6.6}$$

$$h_{CU} = \frac{b_{CU}}{b_Q} = \frac{b_{CU,CU}}{b_{Q,CU}} = \frac{b_{Q,CU}}{b_{Q,Q}} = \frac{b_{CU,AS}}{b_{Q,AS}} = \frac{d_{PL,CU}}{d_{PL,Q}} = \frac{d_{PM,CU}}{d_{PM,Q}}. \tag{6.7}$$

The estimation results and the choice of the most appropriate form will be made in Section 6.4.1, after the description of data that is provided in Section 6.3.

Finally, we have to make a decision on the method used to estimate the cost frontier function. Since the main objective of this work is to measure cost inefficiency, several SFA methods will be applied to estimate the chosen functional form of the specified cost function in Eq.(6.1). The stochastic frontier cost function will be estimated using several of the methods discussed in great detail in Chapter 5. The differences between the various specifications are related to the assumptions imposed on the error term (ε_{it}) introduced in Eq.(6.2), cost inefficiency and firm-specific effects. Six different estimation methods are considered. Table 6.2 summarises the models used in the analysis. The purpose of estimating different models is to check the consistency of the coefficients and efficiency scores

[104] Restrictions are derived by comparing the parameters of Eq.(6.2) and Eq.(6.5).

obtained. In order for the SFA analysis to be applied in a regulation process it has to produce reliable results.

Model I is based on the COLS method which is, in fact, a deterministic frontier method. It does not allow for stochastic errors and it assumes that all deviations of observations from the theoretical minimum are attributed solely to the inefficiency of firms. We can overcome this shortcoming by using stochastic frontier methods. **Model II** is a pooled frontier model estimated by the ML method proposed by Aigner, Lovell and Schmidt (1977). Since it does not assume any firm-specific effects, it does not have the ability to distinguish between the cost inefficiency and unobserved heterogeneity of the firms. This shortcoming can be improved by some of the panel data stochastic frontier models.

Table 6.2: Econometric specification of the models employed

Model	*Firm-specific component* α_i	*Random error* ε_{it}	*Inefficiency* u_{it}
Model I COLS	None	$\text{iid}(0,\sigma_\varepsilon^2)$	$\hat{\varepsilon}_{it} + \lvert\min\{\hat{\varepsilon}_{it}\}\rvert$
Model II Pooled (ML)	None	$\varepsilon_{it} = v_{it} + u_{it}$ $u_{it} \sim \text{iid}N^+(0,\sigma_u^2)$ $v_{it} \sim \text{iid}\,N(0,\sigma_v^2)$	$E(u_{it}\lvert\varepsilon_{it})$
Model III RE (GLS)	$\alpha_i \sim \text{iid}(0,\sigma_\alpha^2)$	$\varepsilon_{it} = v_{it} + \alpha_i$ $\text{iid}(0,\sigma_\varepsilon^2)$	$\hat{\alpha}_i - \min_i\{\hat{\alpha}_i\}$
Model IV RE (ML)	$u_i \sim \text{iid}\,N^+(0,\sigma_u^2)$	$\varepsilon_{it} = v_{it} + u_i$ $v_{it} \sim \text{iid}\,N(0,\sigma_v^2)$	$E(u_i\lvert\varepsilon_i)$
Model V TFE	Fixed (group dummies α_i)	$\varepsilon_{it} = v_{it} + u_{it}$ $u_{it} \sim \text{iid}N^+(0,\sigma_u^2)$ $v_{it} \sim \text{iid}\,N(0,\sigma_v^2)$	$E(u_{it}\lvert\varepsilon_{it})$
Model VI RE (GLS) + Mundlak	$\alpha_i = \alpha_0 + \gamma'\bar{\mathbf{x}}_i + \delta_i$ $\bar{\mathbf{x}}_i = (1/T_i)\sum_1^T \mathbf{x}_{it}$ $\delta_i \sim \text{iid}(0,\sigma_\delta^2)$	$\varepsilon_{it} = v_{it} + \delta_i$ $\text{iid}(0,\sigma_\varepsilon^2)$	$\hat{\delta}_i - \min_i\{\hat{\delta}_i\}$

In **Model III** we first consider the random-effects model proposed by Schmidt and Sickles (1984). The model is estimated by the feasible Generalised Least Squares (GLS) method. If we allow stronger distributional assumptions on the inefficiency term to hold, we can use the ML procedure to estimate the RE model, which will be done in **Model IV**. The latter method was introduced by Pitt and Lee (1981). The FE estimator is considered inappropriate since its

precision relies on the within variation which is very low in our case. The within variation in our sample accounts for just 0.7% of the total variation of the dependent variable total cost. Also, time-invariant variables which are often presented in the network industries cannot be included in the FE model. Nonetheless, the appeal of the FE model as opposed to the GLS estimator is that the former produces unbiased estimates of the regression coefficients even if the firm-specific effects are correlated with the regressors. On the other hand, the FE model produces biased inefficiency estimates due to the incidental parameter problem. Based on the above arguments we decided not to employ this model.

The main weakness of Model III and Model IV is that time-invariant cost inefficiency is assumed.[105] Consequently, these models do not have the ability to distinguish between time-invariant unobserved heterogeneity and cost inefficiency. Any time-invariant firm-specific effects are treated as inefficiency. Therefore, we additionally estimate the stochastic frontier cost function by applying true fixed effects (**Model V**) introduced by Greene (2002a). The true fixed effects (TFE) model treats firm-specific fixed effects and inefficiency separately and is therefore able to distinguish between the unobserved heterogeneity and inefficiency. In this way it tries to overcome some limitations of the conventional fixed effects model. Since the model is estimated by 'brute force' maximum likelihood, its results do not depend on the within variation. The remaining shortcoming of the TFE model is the incidental parameters problem.[106] Further, time-invariant firm characteristics cannot be included in the model as explanatory variables. Nevertheless, these effects are viewed as unobserved heterogeneity and are (at least partially) captured by the firm-specific time-invariant term additionally specified by this model. The true random effects (TRE) model proposed by Greene (2002b) was also applied, but the simulated maximum likelihood estimation method did not converge. Therefore, this method cannot be considered in estimation of the cost function since the obtained results are unreliable. In Section 6.3.2, where the estimation results of different methods are presented, we also provide a possible explanation why the TRE model was not found to perform well in our case. Finally, **Model VI** is used to control for the correlation between unobserved heterogeneity and explanatory variables. It uses Mundlak's (1978) specification of the RE model (Model III). This specification was introduced to stochastic frontier analysis by Farsi, Filippini and Kuenzle (2005). The resulting GLS estimator of the RE model is unbiased and identical to the 'within' estimator

[105] Battese and Coelli's (1992) parameterisation of time effects (time-varying decay model) was applied as well, where the results did not much differ from Model IV. The Battese and Coelli (1992) model was estimated by Stata 8.0.

[106] Greene (2005) finds the bias to be small with respect to the estimates of the regression coefficients. For the inefficiency estimates the bias is found to be larger, where an overestimation error of about 25% is reported.

of the FE model. Therefore, due to the low within variation in our sample the regression coefficients may be imprecisely estimated.

6.3 Data Description

The study is based on a panel data set for Slovenian water distribution utilities over the 1997-2003 period. Since water supply utilities fall within the responsibility of local communities, the data on their operation were not collected systematically in the past. Nevertheless, the Ministry of the Environment and Spatial Planning realised that gathering data at the national level is needed in order to monitor and compare companies' performances. In addition, new legislation on the price regulation of these companies envisages the benchmarking of utilities in the future. Thus, the data were collected from the public utilities via a questionnaire issued by the Ministry of the Environment and Spatial Planning.[107] In this way we obtained data on 52 water supply utilities over the 1997-2003 period. The number of observations for each utility varies from 2 to 7 years; on average, there are 6.38 observations per utility. The sample is an unbalanced panel consisting of a total of 332 observations.

Utilities included in the sample supply 153 out of the 192 municipalities in Slovenia, that is almost 80% of all municipalities. All Slovenian regions are covered by the utilities in the sample. In Appendix II a list of the utilities included in the sample is provided, ranked from the largest to smallest in size with respect to the output delivered. There is also information on the legal form, region and number of municipalities served by the utility. Only four companies in the sample are not public utilities, they operate as private companies or have a concession. The utilities differ in terms of size and as well in some environmental conditions. Some utilities also provide other services like wastewater treatment, solid waste disposal etc. Usually, in smaller municipalities all communal activities are joined within a single company, while in larger municipalities communal activities are provided separately by several companies.

Since 1997, utilities have been obliged to maintain separate accounts for the different regulated activities. This was facilitated by a modification to the Slovenian Accounting Standards ('SRS') which since 1997 have included a special standard (SRS 35) to deal with issues specific to public sector utilities. According to the SRS 35, utilities providing public services are obliged not only to have separate accounts for regulated and unregulated

[107] The questionnaire was prepared by the Faculty of Economics at the University of Ljubljana.

activities but also, within regulated activities, they should have separate accounts for their different regulated activities. The separation of activities is usually carried out in order to avoid cross subsidisation, increase transparency and enable the easier monitoring of regulated utilities. It also facilitates the cost efficiency analysis of each public service activity separately.

In this study the data collected refer to the supply of drinking water only. In this way the comparability of the data is assured. However, since data on other activities is unavailable it is impossible to study the multi-product cost function and economies of scope. Further, this also raises the question of potential data manipulation since it is not evident whether the multi-utility companies use the same methodology to allocate costs between different activities. If some kind of data manipulation is present in the public utilities, this would have a considerable influence on the estimated cost inefficiencies of the water distribution activity. Unfortunately, based on the data available this issue cannot be investigated any further.

The dataset on Slovenian water distribution utilities contains data from income statements, balance sheets, physical quantities, environmental characteristics and technical data referring to the water distribution activity. Descriptive statistics of the variables included in the model are presented in Table 6.3. The total distribution cost (C) equals the operating and capital expenditure. The price of labour (P_L) is equal to the average annual wages, estimated as labour expenditures divided by the average number of employees for a given year. The price of capital (P_K) is calculated as the ratio of capital cost and the capital stock, which is approximated by the capacity of pumps measured in litres per second. Capital cost consists of depreciation and interests. The price of material (P_M) is obtained by dividing material cost by the length of the distribution network in kilometres. Material cost consists of various groups of costs obtained when subtracting capital and labour costs from the total company's costs. Material cost thus includes the cost of energy, material and services. All input prices and costs were deflated to 2000 constant Slovenian tolars (SIT) using the producers' price index and are expressed in thousands of tolars.

The output (Q) is measured as the amount of water supplied to the final customers expressed in cubic metres. The number of final customers (CU) is the sum of household and non-household customers. The size of the service area (AS) is expressed in square kilometres. Water losses are obtained as the difference between the amount of water pumped into the distribution system and the amount of water supplied to final customers. The share of water losses is calculated as the ratio of water losses and the water pumped into the pipes. It is considered that the utility has low water loss levels if the share of water losses does not

exceed the first quartile, which equals 20% of water losses. The variable is included in the model as a dummy variable D_{LOSL} with a value of 1 if the firm has low water losses, and 0 otherwise.[108] In some cases, water needs to be treated in order to be suitable for drinking. A dummy variable D_{TREAT} takes a value of 1 if the firm distributes water that has to be treated chemically before distribution and a 0 value otherwise. Only demanding chemical treatment is taken into the account; simple chemical treatment (disinfection and chlorination) is not considered. Since water distribution utilities can use surface water, underground water or a mix of both resources, the type of water resource is also included in the model. D_S is a dummy variable for the use of surface water only and D_U is a dummy variable for the use of underground water only.

Table 6.3: Descriptive statistics

Variable description	Variable	Mean	Std. Dev.	Minimum	Maximum
Total annual cost (10^3 SIT)[1]	*TOTEX*	304,698.2	538,387.0	7,208.0	2,997,533.8
Price of labour (10^3 SIT/ employee)	*PL*	3,047.7	397.1	2,131.9	4,162.7
Price of capital (10^3 SIT/ litre per sec.)	*PK*	449.4	564.9	13.5	1,484.0
Price of material (10^3 SIT/ km of network)	*PM*	312.0	244.3	46.9	1,412.0
Water supplied (m^3)	*Y*	2,298.780	3,835,452	106,627	25,507,653
Number of customers	*CUST*	7,402.1	7,777.4	515.0	43,272.0
Size of service area (km^2)	*AREA*	336.9	240.0	57.8	949.1
Treatment dummy	D_{TREAT}	0.120	0.326	0	1
Dummy for surface water	D_S	0.199	0.400	0	1
Dummy for underground water	D_U	0.355	0.479	0	1
Dummy for low water losses	D_{LOSL}	0.250	0.434	0	1

[1] The average official exchange rate of Slovenian tolar (SIT) in 2000 was 1 EUR = 205,0316 SIT (Bank of Slovenia, 2001).

[108] We did not include the numerical variable measuring percentage water losses in the model since this would increase the number of coefficients to be estimated. Due to the high correlation between the output, number of customers, size of the service area and water losses, multicollinearity problems could arise. Thus, we decided instead to create the dummy variable for water losses.

6.4 Parameter Estimates of the Cost Frontier Function

6.4.1 Choice of the Functional Form

In Section 6.2 we narrowed the choice of the functional form to choosing between Cobb-Douglas, general translog and hedonic translog functions. Having selected three estimable functional forms with plausible theoretical and applicative properties, we will base our final decision on statistical criteria and data-specific considerations. Accordingly, a likelihood ratio test is performed to determine the functional form that provides the best fit for the data.

The Cobb-Douglas and general translog cost functions are both linear in parameters so the parameters can easily be estimated by using linear least squares techniques. Since the hedonic translog function is non-linear in parameters, a nonlinear maximum likelihood technique has to be employed to estimate the parameters. It should be noted that at this point we are not yet interested in an estimation of cost inefficiencies but rather on determining the functional form that provides the best fit for the data. Therefore, the functional form should be chosen independently of assumptions regarding the distribution of inefficiency. Rather than employing SFA methods at this stage, the OLS method is used to obtain preliminary parameter estimates of the Cobb-Douglas and the translog cost function, and the Nonlinear Least Squares (NLS) method is used for the hedonic translog function. The estimated parameters of the Cobb-Douglas, the general translog and the hedonic translog cost function are given in Appendix III, Table III.1. The Cobb Douglas and translog functions were also estimated by the GLS estimator, which led us to the same conclusion with respect to the preferable functional form as in the OLS case. Since the panel data structure is not taken into account in the NLS estimation of the hedonic cost function, we compare its estimates with the OLS estimates of the Cobb-Douglas and translog functions. It can also be noted that the time trend t is not included in the estimation because the choice of the functional form should be driven by the data and not by the assumption regarding technical change. However, including the time trend does not significantly change the results.

To compare the translog specification against Cobb-Douglas and, alternatively, against the hedonic translog specification, the likelihood ratio (LR) test is used. The null hypothesis of the LR test imposes restrictions on the parameters of unrestricted model, in our case the general translog. If the restrictions are valid, then imposing them should not lead to a large reduction in the log-likelihood function. The likelihood ratio equals $\lambda = L_R/L_U$, where L_R and L_U are the likelihood functions of the restricted and unrestricted models evaluated at the estimated vector of the parameters of the respective model. The limiting distribution of –2ln

λ is χ^2, with degrees of freedom equal to the number of restrictions imposed (Greene, 2000).[109] The respective chi-square statistics resulting from the comparison of alternative functional forms for our model are reported in Appendix III, Table III.2. The chi-square statistics are highly significant in both cases, i.e. translog vs. Cobb-Douglas and translog vs. hedonic, indicating that restrictions in the null hypothesis should be rejected. This indicates that the general translog functional form is the most appropriate one and should therefore be applied to estimate the cost function. Since both the Cobb-Douglas and hedonic functional forms impose certain restrictive assumptions, the locally flexible translog specification is also the most appropriate form from the theoretical point of view. Hence, the translog functional form is utilised to estimate the total cost function of Slovenian water distribution utilities.

6.4.2 Estimation Results

The estimation results of the translog cost frontier function of Slovenian water distribution utilities obtained by the six different models are given in Table 6.4.[110] The expansion point of the stochastic frontier cost function specified in Eq.(6.2) is chosen to be the sample median, i.e. the median values of variables included in the model. Since total cost and all the continuous explanatory variables are in logarithms and normalised by their medians, the estimated first-order coefficients can be interpreted as cost elasticities evaluated at the sample median.

In terms of the coefficients' significance and expected coefficients' sign, Model I (the COLS model), Model II (the Pooled ML model) and Model V (the 'true' FE model) seem to perform better than Model III (the RE model estimated by the GLS technique), Model IV (the RE model estimated by the ML estimation procedure) and Model VI (Mundlak's formulation of the RE model estimated by GLS). In particular, the coefficients of Model VI are mostly insignificant, likely due to the large number of explanatory variables included in the model and the high correlation between them.

[109] In our case, 15 restrictions had to be imposed on the translog to obtain the Cobb-Douglas specification, and 12 restrictions for the hedonic specification.

[110] The models are estimated using NLogit 3.0.

Table 6.4: Estimation results of the frontier cost function

Coefficient	*Model I* COLS	*Model II* Pooled (ML)	*Model III* RE (GLS)	*Model IV* RE (ML)	*Model V* TFE	*Model VI* RE (GLS) + Mundlak
$\ln a$	11.321***	11.570***	11.856***	11.424***	-	11.816***
	(0.037)	(0.036)	(0.054)	(0.079)		(0.069)
c_{PL}	0.585***	0.579***	0.405***	0.401***	0.521***	0.374***
	(0.024)	(0.024)	(0.024)	(0.041)	(0.030)	(0.029)
c_{PM}	0.188***	0.180***	0.341***	0.339***	0.193***	0.354***
	(0.023)	(0.022)	(0.020)	(0.042)	(0.028)	(0.025)
b_Q	0.361***	0.329***	0.289***	0.290***	0.258***	0.157*
	(0.063)	(0.059)	(0.063)	(0.091)	(0.069)	(0.092)
b_{CU}	0.433***	0.449***	0.471***	0.454***	0.503***	0.298*
	(0.063)	(0.059)	(0.071)	(0.095)	(0.072)	(0.168)
b_{AS}	0.172***	0.193***	0.201***	0.218***	0.158***	0.200
	(0.024)	(0.023)	(0.040)	(0.074)	(0.032)	(0.132)
$c_{PL,PL}$	-0.110**	-0.097**	0.034	0.015	-0.178***	0.070*
	(0.051)	(0.047)	(0.037)	(0.069)	(0.060)	(0.042)
$c_{PM,PM}$	-0.084*	-0.052	0.014	-0.005	-0.109**	0.033
	(0.038)	(0.036)	(0.028)	(0.044)	(0.046)	(0.032)
$c_{PL,PM}$	0.159***	0.144***	0.023	0.040	0.222***	-0.010
	(0.041)	(0.037)	(0.029)	(0.050)	(0.050)	(0.033)
$b_{Q,Q}$	0.426**	0.587***	0.330**	0.248	0.673***	0.176
	(0.149)	(0.152)	(0.124)	(0.239)	(0.149)	(0.164)
$b_{CU,CU}$	0.010	0.122	-0.029	-0.094	-0.184	-0.051
	(0.236)	(0.226)	(0.189)	(0.281)	(0.252)	(0.258)
$b_{AS,AS}$	0.122*	0.195***	0.086	0.026	0.287***	0.673
	(0.063)	(0.055)	(0.116)	(0.163)	(0.078)	(0.973)
$b_{Q,CU}$	-0.301*	-0.432**	-0.209	-0.149	-0.350*	-0.138
	(0.179)	(0.177)	(0.139)	(0.228)	(0.183)	(0.164)
$b_{Q,AS}$	0.009	0.022	-0.032	-0.031	-0.103	-0.058
	(0.083)	(0.076)	(0.088)	(0.121)	(0.089)	(0.128)
$b_{CU,AS}$	0.173**	0.155**	0.127	0.185	0.250***	-0.024
	(0.082)	(0.078)	(0.092)	(0.174)	(0.093)	(0.259)
$d_{PL,Q}$	0.097	0.096	0.104	0.117	0.089	0.112
	(0.081)	(0.080)	(0.063)	(0.118)	(0.089)	(0.074)
$d_{PL,CU}$	-0.107	-0.056	-0.137**	-0.157	-0.060	-0.157*
	(0.079)	(0.078)	(0.066)	(0.119)	(0.093)	(0.077)
$d_{PL,AS}$	-0.014	-0.065*	0.031	0.028	-0.105**	0.060
	(0.039)	(0.037)	(0.039)	(0.076)	(0.048)	(0.049)
$d_{PM,Q}$	-0.078	-0.109*	-0.101**	-0.092	-0.148**	-0.074
	(0.061)	(0.059)	(0.050)	(0.069)	(0.069)	(0.059)
$d_{PM,CU}$	0.092	0.094	0.126**	0.135*	0.093	0.107*
	(0.066)	(0.065)	(0.051)	(0.067)	(0.079)	(0.059)
$d_{PM,AS}$	0.023	0.046	-0.055*	-0.081	0.124***	-0.095**
	(0.037)	(0.035)	(0.030)	(0.081)	(0.044)	(0.036)
h_T	-0.001	0.002	-0.002	-0.002	-0.008*	0.001
	(0.005)	(0.005)	(0.003)	(0.004)	(0.004)	(0.003)
g_S	0.176***	0.202***	0.097	0.216*	-	0.209***
	(0.033)	(0.029)	(0.069)	(0.116)		(0.074)
g_U	0.057**	0.090***	0.040	0.219*	-	0.059
	(0.028)	(0.026)	(0.059)	(0.201)		(0.064)
g_{TREAT}	0.117***	0.120***	0.212***	0.287*	-	0.082*
	(0.037)	(0.037)	(0.080)	(0.164)		(0.088)
g_{LOSL}	-0.175***	-0.156***	-0.037	-0.020	-	-0.042*
	(0.029)	(0.027)	(0.020)	(0.022)		(0.022)

Table 6.4: Continuation

Coefficient	*Model I* COLS	*Model II* Pooled (ML)	*Model III* RE (GLS)	*Model IV* RE (ML)	*Model V* TFE	*Model VI* RE (GLS) + Mundlak
a_{PL}	-	-	-	-	-	0.185** (0.066)
a_{PM}	-	-	-	-	-	-0.170** (0.064)
a_{Q}	-	-	-	-	-	0.158 (0.190)
a_{CU}	-	-	-	-	-	0.201 (0.232)
a_{AS}	-	-	-	-	-	-0.049 (0.145)
$a_{PL,PL}$	-	-	-	-	-	-0.244 (0.173)
$a_{PM,PM}$	-	-	-	-	-	-0.189 (0.118)
$a_{Q,Q}$	-	-	-	-	-	0.258 (0.461)
$a_{CU,CU}$	-	-	-	-	-	-0.132 (0.711)
$a_{AS,AS}$	-	-	-	-	-	-0.479 (0.986)
$a_{PL,PM}$	-	-	-	-	-	0.242* (0.132)
$a_{PL,Q}$	-	-	-	-	-	0.081 (0.257)
$a_{PL,CU}$	-	-	-	-	-	0.007 (0.239)
$a_{PL,AS}$	-	-	-	-	-	-0.106 (0.113)
$a_{PM,Q}$	-	-	-	-	-	-0.056 (0.178)
$a_{PM,CU}$	-	-	-	-	-	-0.033 (0.197)
$a_{PM,AS}$	-	-	-	-	-	0.201* (0.112)
$a_{Q,CU}$	-	-	-	-	-	-0.074 (0.543)
$a_{Q,AS}$	-	-	-	-	-	0.006 (0.245)
$a_{CU,AS}$	-	-	-	-	-	0.238 (0.333)
σ_v (s_v)	0.1856	0.0976	0.0712	0.0698	0.1542	0.0741
σ_u (s_u)	-	0.2502	0.1714	0.4282	0.2611	0.1616
$\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$	-	0.2686****	-	0.4338****	0.3032****	-
$\lambda = \sigma_u / \sigma_v$	-	2.564**** (0.3397)	-	6.137** (3.0475)	1.693**** (0.2079)	-

Notes: standard errors in brackets;

* – significant at 10%, ** – significant at 5%, *** – significant at 1%, **** – significant at 0.1% (two-sided significance level)

As mentioned, we also tried to apply the true random effects model but the estimation method did not converge. The TRE estimates are provided in Appendix IV. The results of the TRE model are in line with the reported results from other models, however, they cannot be regarded as reliable since the method did not converge. A possible explanation why this model did not perform well is that the model specification is too rich for our data and, as a result, some of the error terms degenerate to zero. In our case, this happens to the error term u_{it} that is supposed to capture the cost inefficiency (σ_u is not found to be significantly different from zero). It seems that all effects are already captured by the random error v_i and the firm-specific term ω_i (σ_v and σ_ω are found to be significant), so there is nothing left to be captured by the inefficiency term u_{it}. Recall that in the TRE model the time-invariant firm-specific effect ω_i is additionally introduced in the model to capture unobserved heterogeneity.

In all other ML models (Model II, IV and V) the inefficiency term is found to be significant, as confirmed by the λ statistic that compares σ_u and σ_v (see Table 6.3). By comparing the estimated variance of the random error (s_v) and firm-specific effects (s_u) in Models III and VI we can also confirm the relative importance of firm-specific effects that are supposed to capture cost inefficiency. We turn to the analysis of inefficiency estimates in the next chapter.

The results of the six models show that the output coefficient (b_Q) is positive and highly significant in all models. It suggests that, on average, a one percent increase in the amount of water supplied will increase the total cost of Slovenian water distribution utilities by 0.25% to 0.36%, depending on the model considered. Model VI produces a much lower estimate, amounting to 0.157. A possible explanation is that the resulting estimator of Model VI equals the within estimator and, due to the very low within variation in our sample, the results of this model may be imprecise. Similarly, the coefficients of the two output characteristics, the number of customers (b_{CU}) and the size of service area (b_{AS}), are found to be significantly positive. The coefficient of the number of customers varies between 0.43 and 0.50, with the exception of Model VI where the coefficient is again much lower. The coefficient of the service area size is found to be between 0.16 and 0.22. In Model VI, b_{AS} is not found to be significant. The sum of the three coefficients (b_Q, b_{CU} and b_{AS}) at the sample median does not exceed 1, which will prove to be important when analysing the economies of scale of Slovenian water distribution utilities.

With respect to input prices, the cost function has to satisfy the properties of being non-decreasing, linearly homogeneous and concave in input prices. The property of linear homogeneity in input prices holds since it was imposed prior to the estimation. The cost

function is non-decreasing in input prices since both the labour price coefficient as well as the material price coefficient are positive and highly significant. The estimated coefficient for the labour price (c_{PL}), representing the share of costs attributed to labour at the median point, amounts to around 0.4 in the two RE models (Models III and IV), more than 0.5 in the TFE model (Model V) and slightly less than 0.6 in Model I and Model II, where the last two models do not take the panel aspect of the data into account. The estimated coefficient for the material price (c_{PM}) is found to be less than 0.2 in Models I, II and V and around 0.34 in Models III and IV. Again, in Model VI the estimated coefficients c_{PL} and c_{PM} are much lower. At the same time, the significant coefficients a_{PL} and a_{PM} indicate the presence of a correlation between the explanatory variables and unobserved heterogeneity captured by individual effects in the RE models. Model VI controls for this heterogeneity and produces unbiased results given enough within variation, which is not the case in our study. The sum of the two coefficients (c_{PL} and c_{PM}) in the six models varies from 0.71 to 0.77 implying that the share of the capital cost is estimated to be between 0.23 and 0.29.[111] Thus, the variation of the share of capital cost across the models is lower than the variation of the respective shares for labour and material.

The concavity in input prices and twice-continuous differentiability of the cost function imply that the Hessian is negative semidefinite. The elements of the Hessian also represent conditional input demand responses to changes in input prices. The coefficients $c_{PL,PL}$, $c_{PM,PM}$, and $c_{PL,PM}$, are only significant in Models I, II and V. In these models they also hold the expected sign. The coefficients $c_{PL,PL}$ and $c_{PM,PM}$, have the negative sign, implying that as labour or material becomes more expensive you buy less of that input. The positive sign of cross-product $c_{PL,PM}$ indicates that as labour (material) becomes more expensive you buy more material (labour).[112] The negative semidefiniteness of the Hessian, comprising second-order derivatives of the cost function with respect to all three input prices (labour, material and capital), is confirmed at the sample median for Models I, II and V. In other models, the respective coefficients are not found to be significant.

Time does not seem to have a significant influence on the costs of Slovenian water distribution utilities. By assuming a one-sided hypothesis, only in Model V are costs found to be significantly decreasing over the analysed period. Based on the results it cannot be

[111] Due to the imposed linear homogeneity, the capital price coefficient is obtained as: $c_{PK}=1-c_{PL}-c_{PM}$. The observed data on shares of labour and material correspond more closely to the two RE models, while the data on capital share closely correspond to the estimate obtained from the TFE model.

[112] From the estimated frontier cost function we can also derive the respective coefficients for capital prices ($c_{PK,PK}$, $c_{PK,PL}$ and $c_{PK,PM}$), which for Models I, II and V also have the expected signs.

concluded that total cost has considerably changed over time. This is largely consistent with non-competitive environment in which the public utilities operate. Also, price regulation is not designed in a way that would stimulate utilities to decrease their costs and operate more efficiently.

On the other hand, some models show that the dummy variables relating to the water resource used, water losses and level of water treatment can significantly influence the cost. Since these variables are time-invariant they had to be omitted from Model V. Nevertheless, they are captured by the firm-specific time invariant term additionally introduced in Model V to capture the effect of unobserved heterogeneity. Using only surface water (g_S) or only underground water (g_U) increases the costs compared to the use of a water resource mix. These two dummies are shown to be significant in Model I and Model II. The coefficient g_S is also significant in Model VI and, if we make a one-sided hypothesis, g_S is in addition significant in Model IV. Further, the use of heavy chemical treatment (g_{TREAT}) significantly increases the costs, with this holding for Models I – IV. Again, in Model IV we have to assume a one-sided test. Finally, low water losses (g_{LOSL}) significantly decrease total costs in Model I and Model II and, with a one-sided test, also in Models III and VI.

After estimating the different models, their performance is usually evaluated and the choice of the preferred model is made. The decision has to consider the performance of the analysed models with respect to the coefficient estimates as well as the inefficiency estimates. Hence, before opening this discussion the inefficiency scores obtained by the different models have to be analysed. This is done in the following chapter.

7 Cost Inefficiency, Economies of Scale and Alternative Uses of the Estimated Cost (Frontier) Function

In this chapter, the cost inefficiency of the utilities is estimated based on the results obtained in the previous chapter. The consistency of the inefficiency estimates resulting from the different models is tested. Estimates of economies of output density, customer density and economies of scale will also be obtained from the estimated cost frontier function. In addition, alternative uses of the cost (frontier) function will be explored, namely for cost predictions and for the decomposition of total factor productivity growth. TFP growth as another important measure of firm performance is analysed. We as well consider possible implications of the obtained econometric results for economic policy-making and discuss in what way the results could be used in the price regulation of Slovenian water utilities.

7.1 Estimated Inefficiency Scores and Their Consistency

Table 7.1 provides descriptive statistics on the cost inefficiency estimates of Slovenian water distribution utilities obtained from Models I – VI. We can observe some notable differences in the estimated cost inefficiency levels.[113] As expected, the average estimated cost inefficiency of 70.4% is the highest in the COLS model (Model I) since this method does not allow for the random error and attributes all deviations from the frontier to inefficiency. By employing the pooled stochastic frontier model (Model II), the average cost inefficiency drops considerably and is estimated to be 22.5%. The shortcoming of this model (and of Model I) is that it does not take into account the panel aspect of the data. In the case of the RE panel data stochastic frontier models, the estimated average cost inefficiencies are again quite high; the inefficiency amounts to 66.3% in the RE GLS model (Model III) and 50% in the RE ML model (Model IV). By applying the Mundlak formulation of the RE GLS model and thus avoiding possible problems resulting from correlation between firm-specific effects and explanatory variables, the estimated inefficiency drops to 43.4% (Model VI). The relatively high inefficiency levels of the RE models might to some extent be attributed to unobserved firm-specific time-invariant effects. The RE models treat these effects as time-

[113] A cost-efficiency score (EF_i) can be obtained as the inverse of a cost-inefficiency score (EFF_i). The value $1 - EF_i$ represents the reduction in total costs needed to achieve the minimum efficient cost level.

invariant cost inefficiency so the cost inefficiency estimates obtained by these models are most likely overestimated. This is not the case of the pooled model since each observation is treated as independent and, accordingly, the inefficiency is considered to vary across utilities and over time. It is thus unlikely that the inefficiency term in this model would capture time-invariant firm-specific effects. Further, in the RE models the median values of cost inefficiency are considerably lower compared to the means, indicating that the means are influenced by the extreme values.

Table 7.1: Estimated cost inefficiency scores

Inefficiency score (EFF_i)	*Model I* COLS	*Model II* Pooled (ML)	*Model III* RE (GLS)	*Model IV* RE (ML)	*Model V* TFE	*Model VI* RE (GLS) + Mundlak
Mean	1.704	1.225	1.663	1.500	1.191	1.434
Median	1.660	1.181	1.556	1.378	1.182	1.378
Std. Dev.	0.319	0.162	0.376	0.346	0.057	0.242
Minimum	1.000	1.031	1.000	1.118	1.067	1.000
Maximum	3.146	1.710	2.690	2.599	1.514	2.142

Finally, the average cost inefficiency based on the true fixed effects model (Model V) is estimated to be 19.1%, which is in line with Model II. Relatively low inefficiency estimates are expected since the true fixed effects model (contrary to the RE models considered) is able to distinguish unobserved firm-specific fixed effects from inefficiency and is thus able to treat the two effects separately. However, one cannot be certain whether the time-invariant effects belong to the unobserved heterogeneity or to the cost inefficiency. The choice of appropriate model is also based on our belief whether some time-invariant unobserved heterogeneity exists in the model or whether inefficiency does not in fact vary over time. In the latter case, the inefficiency scores obtained by the TFE model could in fact be underestimated. Taking into account the non-competitive environment in which Slovenian water distribution utilities operate, the cost inefficiency levels estimated by the TFE model are probably slightly underestimated. Since companies were not obliged to decrease costs and improve efficiency in the analysed period, at least some time-invariant cost inefficiency is expected to be present in these companies. However, remember that the dummy variables of water source used, water losses and level of water treatment could not be included in this model. It is thus reasonable to believe that time-invariant unobserved heterogeneity is present in the model and that the firm-specific time-invariant term mainly captures these effects rather than inefficiency. Further, inefficiency estimates according to the TFE model are found to closely correspond to the pooled model. It may thus be concluded that these two

models set the lower bound for the cost inefficiency of Slovenian water distribution utilities. The actual cost inefficiency is probably slightly higher than the estimates indicated by these two models. Conversely, the RE models are found to largely overestimate the cost inefficiency.

We can now turn to selecting the most appropriate model. The choice has to take into account the performance of the model with respect to both the coefficient estimates and the inefficiency estimates. Conventional panel data models were primarily designed to estimate coefficients of a given function. Theoretical findings imply that, given sufficient within variation, the FE model produces unbiased coefficient estimates while on the other hand the RE estimator is found to be more efficient than the FE estimator given that firm-specific effects are uncorrelated with the random error and with the regressors. These two models were, however, not originally designed to estimate inefficiency. It is therefore not surprising that they possess several shortcomings with respect to estimating inefficiency. In contrast, stochastic frontier models were primarily constructed for modelling the inefficiency. Their primary interest is the error term and its structure and not the coefficients of the specified function so they are generally found to perform better with respect to the former issue. Hence, there is a trade-off present between performing well as regards the coefficient estimates on one side and performing well with respect to the inefficiency estimates on the other. Nevertheless, based on the coefficient estimates in Table 6.4 in our case pooled model and TFE model are found to perform better than other models in terms of the coefficients' significance and expected coefficients' sign. These two models are also found to be the preferred models in a cost inefficiency estimation since in the other models a large part of unobserved heterogeneity is mistakenly treated as inefficiency. Since the TFE model also takes the firm-specific effects into account and explicitly deals with the problem of separating unobserved heterogeneity and inefficiency, it is chosen as the preferred model.

To be able to reliably use stochastic frontier methods in the price regulation of utilities, different methods should provide similar results regarding the utilities' inefficiency scores and rankings. Therefore, it is important to check the consistency of the inefficiency results obtained. If consistency is not established, the regulator cannot directly use inefficiency estimates to set requirements for cost reductions but can merely use the results to determine the range in which the inefficiency scores of the utilities may be located. In this case, SFA can only be used as a complementary instrument in the price-regulation process. Bauer et al. (1998) proposed a set of consistency conditions which frontier efficiency measures should meet so as to be most useful for regulatory purposes. The consistency conditions are:

(i) the efficiency scores generated by the different approaches should have comparable means, standard deviations and other distributional properties;

(ii) the different approaches should rank the companies in approximately the same order;
(iii) the different approaches should identify mostly the same companies as the best practice and the worst practice;
(iv) the different approaches should demonstrate reasonable stability over time (i.e., tend to identify the same companies as relatively efficient and inefficient in different years, rather than varying markedly from one year to the next);
(v) the efficiency scores generated by different approaches should be reasonably consistent with competitive conditions in the market; and
(vi) the efficiencies from the different approaches should be reasonably consistent with standard non-frontier performance measures, such as return on assets or the cost/revenue ratio.

Consistency conditions (i), (ii) and (iii) measure the degree to which the different methods are mutually consistent, while conditions (iv), (v) and (vi) measure the degree to which the efficiencies generated by the different models are consistent with reality (Bauer et al., 1998).

Descriptive statistics of the cost inefficiency estimates obtained by the six different models have already been provided. Based on the results reported in Table 7.1 we established notable differences in the cost inefficiency levels and provided a possible explanation for these differences. The three random effects models (Model III, IV and VI) resulted in comparable mean efficiency levels which are, as expected, lower than the average inefficiency level of the COLS model and significantly higher than the average inefficiency levels of the pooled and TFE models. We can also notice differences in standard deviations of the inefficiency scores. Again, the three RE models as well as the COLS model produce comparable standard deviations of the inefficiency scores, which are considerably higher than in the pooled and the TFE models. Similar conclusions can be drawn with respect to the estimated maximum cost inefficiency and the range of variation of inefficiency scores.

In Appendix V (Figure V.1 – 6), the distributions of the inefficiency scores resulting from Models I – VI as represented by the kernel distribution functions are given.[114] Except for the COLS method, where the inefficiency term is by construction normally distributed, it can be noticed that the inefficiency terms are positively skewed. In fact, the ML methods need

[114] The Kernel density estimator is a useful substitute for the histogram as a descriptive tool for the underlying distribution that produced a sample of data. Particularly for small samples and widely dispersed data, histograms tend to be 'rough' and difficult to make informative. Thus, the kernel density estimator can be employed as a device to describe the distribution of a variable nonparametrically, that is, without any assumption of the underlying distribution (Greene, 2000).

skewed errors in order for the SFA model to be computable, while the GLS estimator can be estimated even in the presence of non-skewed errors. Once again, the resulting distributions of the inefficiency estimates obtained by the RE models are relatively similar. Compared to the RE models, the range in which the inefficiency estimates can be found is relatively narrow in the pooled model case and particularly narrow in the TFE model. We can also observe that the MLE estimation techniques (Models II, IV and V) produce smoother kernel density functions with fewer irregularities than the GLS methods (Models III and VI). To test for the equality of the cost inefficiency distributions pair-wise, the Kolmogorov-Smirnov test (K–S test) is used. The statistics from a two-sample Kolmogorov-Smirnov equality-of-distributions test are given in Table 7.2. Based on the reported results the null hypothesis of equal distributions is rejected. It can be concluded that the six models we considered all produce different distributions of cost inefficiency scores. Nevertheless, the consistency conditions only require similar and not the same distributions of inefficiency scores.

Table 7.2: Kolmogorov-Smirnov equality-of-distributions test (K–S test)

K-S test (D)[1]	*Model II*	*Model III*	*Model IV*	*Model V*	*Model VI*
Model I	0.7169**	0.1869**	0.4578**	0.8916**	0.4458**
Model II		0.6777**	0.5000**	0.2741**	0.5361**
Model III			0.3343**	0.8554**	0.3434**
Model IV				0.6717**	0.1325 *
Model V					0.7108**

Notes: * – significant at 1%; ** – significant at 0.1%; (two-sided significance level);
[1] The combined K-S test is based on the biggest absolute difference between the inefficiency estimates from the two distributions.

What remains to be tested is whether the models provide similar rankings of the utilities with respect to the cost inefficiency scores. From the regulatory point of view, this issue is considered to be vital. Table 7.3 provides the pair-wise Pearson correlation coefficients between the cost inefficiency estimates. We can observe that, with the exception of Model V, the correlation between the inefficiency scores resulting from different models is positive, significant and, overall, not particularly high. The correlation is especially high between the inefficiency scores from Model I and Model II (non-panel data models), and between the inefficiency scores from Model III and Model IV (RE panel data models). The correlation between the inefficiency scores from Model VI and those from Models I – IV is also relatively high. The correlation between inefficiency scores from Model V and Model I or Model II is significant but quite moderate, whereas the correlation between Model V and

Models III, IV or VI is not significantly different from zero. Again, the reason may be found in the fact that the TFE model treats firm-specific fixed effects (α_i) separately from the inefficiency (u_{it}). Thus, some effects that might be attributed to inefficiency by other models are here captured by the firm-specific effects and thus attributed to firm heterogeneity rather than inefficiency. This may be a plausible reason for the no correlation with all three RE panel data models.

Table 7.3: Correlation between inefficiency scores (Pearson's correlation coefficients)

R	***Model I*** **COLS**	***Model II*** **Pooled**	***Model III*** **RE (GLS)**	***Model IV*** **RE (ML)**	***Model V*** **TFE**	***Model VI*** **RE (GLS) + Mundlak**
Model I	1	0.956*	0.694*	0.627*	0.434*	0.827*
Model II		1	0.667*	0.614*	0.399*	0.838*
Model III			1	0.932*	0.023	0.767*
Model IV				1	0.027	0.696*
Model V					1	0.037
Model VI						1

Note: * – significant at 0.1% (two-sided significance level)

The conclusions based on the rank correlation between the inefficiency scores from different models (Spearman correlation coefficients) are very similar to those found in Table 7.3. With respect to identifying the same best and worst practices for Slovenian water distribution companies, again all models but Model V identify the same (group of) companies as being the most or the least efficient. In contrast, companies identified as the best by Model V are not performing particularly well in the other models. This indicates that the models do not only differ in their estimated inefficiency levels but also in their ranking of the companies. There is more consensus when identifying the worst practice – the worst companies identified by Model V are also not performing that well in the other models.

Figure 7.1 demonstrates that the inefficiency scores of Slovenian water distribution utilities are relatively stable over time. In Models III, IV and VI the inefficiency estimates are constant by construction; they vary only due to the different number of observations across the years. In Models I, II and V the inefficiency estimates are also relatively stable. This is more or less in line with the non-competitive environment in which these utilities operate. Also, there were no regulatory changes during the examined period that would provide incentives for the more efficient production of local monopolies supplying water to final customers.

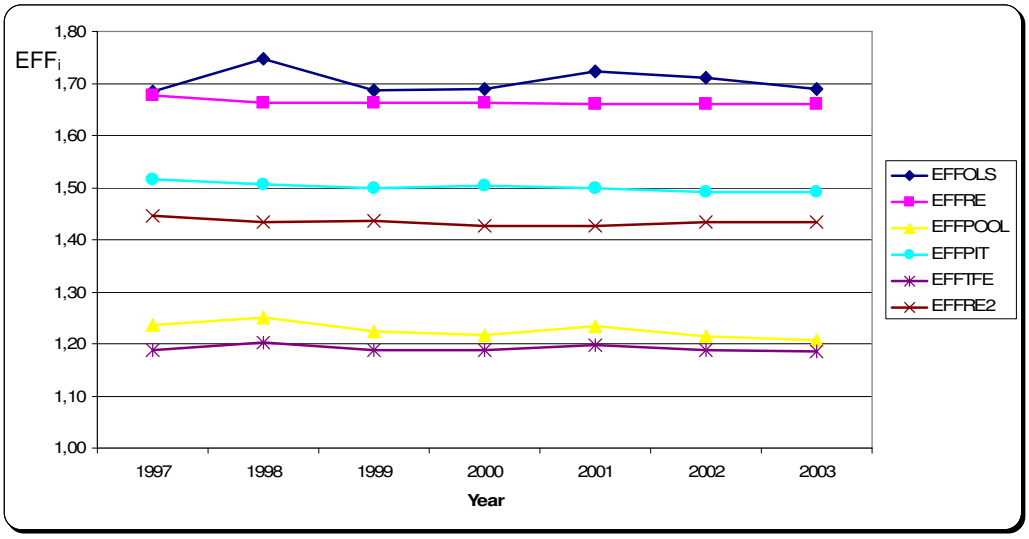

Figure 7.1: Estimated average inefficiency scores by years

Based on the above discussion and interpretation of the results relating to cost inefficiencies it can be concluded that the mutual consistency conditions proposed by Bauer et al. (1998) are not satisfied.[115] These results show the sensitivity of the frontier benchmarking methods in our sample. This is not particularly encouraging since the results cannot be considered as reliable, especially if they are to be applied in the price-regulation process. Therefore, the direct use of inefficiency estimates in the regulation of water distribution utilities may be misleading. Nevertheless, some inconsistency of inefficiency estimates is expected since the various models employ different assumptions regarding cost inefficiency and heterogeneity. We thus cannot expect the results to be completely invariant to these different assumptions.

Whether time-invariant effects belong to unobserved heterogeneity or cost inefficiency is debatable. If there is some time-invariant inefficiency, the inefficiency scores obtained by TFE model could be underestimated. On the other hand, if there is some unobserved time-invariant heterogeneity present the other panel data models treat it as cost inefficiency and thus tend to overestimate it. How we handle time-invariant effects obviously has a large influence on the findings. Ultimately, firm-specific heterogeneity and inefficiency both might contain time-invariant and time-varying elements and there is no perfect way to disentangle them based on the observed data (Greene, 2002a, b). However, we believe that

[115] For example, the inconsistency of inefficiency scores obtained from different models is also established by Farsi and Filippini (2004).

the results obtained by the TFE model can be regarded as a good approximation of the actual cost inefficiency of Slovenian water distribution utilities. Of course, the mechanical use of these results in the price-regulation process is not recommendable. In this case, SFA as a benchmarking tool can only be used as a complementary instrument when regulating prices.

7.2 Prediction Errors

In this section we look at another possibility for employing benchmarking in the price regulation of network utilities. As proposed by Farsi and Filippini (2004), the estimated cost function can also be used to predict the costs of individual companies. Based on this, the regulator can construct confidence intervals for the costs of the companies. In fact, this approach reflects the idea of yardstick competition originally proposed by Schleifer (1985). In his paper it is shown that the yardstick competition concept can be applied to firms producing heterogeneous outputs if these outputs only differ in observable characteristics. To correct for observed heterogeneity, the cost function is used whereby the observable characteristics are included in the cost function in order to correct for cost differences that occur merely due to the heterogeneity of output. The regulator can then use the estimates of the cost function to set corrected yardstick prices for individual firms.[116]

In order to be able to reliably use the above proposed approach for the purpose of price regulation, it has to be confirmed that the model predicts costs with sufficient accuracy. Therefore, the predictive power of the model has to be analysed. To obtain the prediction error of our model, the predicted total costs are compared with the observed total cost. The relative prediction error is then defined as $\hat{\varepsilon}_R = (C_{it} - \hat{C}_{it}) / C_{it}$, where C_{it} is the actual total cost and $\hat{C}_{it}$ is the predicted total cost of the regression.

To estimate the cost function the GLS estimator is chosen. Several arguments can be given in support of this decision. As already pointed out, the FE estimator is unbiased however its precision relies on the within variation. Due to the very low within variation of the total costs in our case, the FE estimator is not considered appropriate. Further, time-invariant variables cannot be included in the FE model so the heterogeneity cannot be fully captured by this model. An advantage of the GLS estimator compared to the SFA models based on the ML estimator is that the former imposes less distributional assumptions on the error term. The only assumption required by the GLS estimator is that the firm-specific effects are not

[116] Recall the discussion in Section 2.2.3.

correlated with the explanatory variables and random error. Moreover, the GLS estimator does not require the composed error term to be positively skewed whereas the ML methods do. Since the main focus here is no longer on the error term structure and inefficiency we are unwilling to make these additional assumptions. In addition, employing the yardstick concept essentially implies that we are reverting from frontier benchmarking to average benchmarking which also has to be reflected in the chosen estimation method.

After estimating the total cost function, the in-the-sample relative prediction errors are calculated. Then we proceed in two directions: (i) by obtaining out-of-sample predictions of costs; and (ii) by forecasting costs. The out-of-sample prediction of the costs involves predicting the costs of a given firm using the estimations obtained from the sample consisting of other firms. Forecasting, on the other hand, involves the prediction of costs in a given year using estimations based on the data previous to this year. One-, two- and three-year-ahead forecasts are considered. The results are summarised in Table 7.4.

As expected, out-of-sample prediction errors and forecast errors are higher than the in-sample prediction errors. From a practical point of view, the results are generally within an acceptable range. The average prediction bias is particularly low for in-sample and out-of-sample prediction errors, 0.2% and -0.3%, respectively. In absolute terms, the average prediction bias is somewhat higher in the case of forecasted errors. A negative sign implies that the forecasted costs tend to be slightly overestimated.

Table 7.4: Relative prediction errors of the RE model (in percent)

Prediction error (GLS estimator)	In-the-sample	Out-of-sample	1 year ahead	2 years ahead	3 years ahead
Average error (absolute value)	4.977	5.860	5.887	7.511	7.174
Average error / prediction bias	0.210	-0.317	-2.162	-2.320	-0.740
Standard Deviation	6.537	8.050	7.294	9.649	9.568
Minimum	-21.017	-39.647	-24.385	-28.097	-30.165
Maximum	19.442	24.512	12.310	17.529	19.848
90^{th} percentile (absolute value)	10.948	13.205	11.075	12.464	12.895
No. of predictions	332	332	52	52	52

The average of the absolute predicted errors is slightly less than 5% for in-sample predictions and somewhat less than 6% for the out-of-sample predictions and 1-year-ahead

forecasts. For the two- and three-year-ahead forecast the respective value is more than 7%, the latter value being surprisingly lower. For 90% of the companies the absolute value of the predicted error is limited to 11% for in-sample predictions, to 13.2% for out-of-sample predictions, and to 11.1%, 12.5% and 12.9% for the one-, two- and three-year-ahead forecasts, respectively.

These results suggest that the random effects panel data model can predict individual total costs with reasonable precision. The regulator could therefore use this model to predict confidence intervals for the utilities' costs. Using such predictions, the regulator could hold utilities within a reasonably well-predicted range of cost efficiency. This approach can in essence be viewed as the rate-of-return regulation combined with benchmarking. The regulator could also deal with the problem of those models with somewhat weaker predictive power by allowing regulated utilities to renegotiate the prices. In this case, the utilities are expected to present credible evidence that would explain why the actual costs are higher than predicted.

7.3 Economies of Scale and Density

In this section we turn to the issue of economies of scale, which can also be an important source of cost reductions in Slovenian water distribution utilities. In addition to the output distributed, several output characteristics can influence the cost of network industries. The inclusion of the number of customers and the size of service area in the cost function allows us to distinguish between economies of output density, economies of customer density and economies of scale (more accurately, economies of size).

From the translog function specified in Eq.(6.2), economies of output density are obtained as follows:

$$E_{OD} = \left(\frac{\partial \ln C}{\partial \ln Q}\right)^{-1} = \left(b_Q + b_{Q,Q} \ln Q + b_{Q,CU} \ln CU + b_{Q,AS} \ln AS + d_{PL,Q} \ln PL^* + d_{PM,Q} \ln PM^*\right)^{-1}. \quad (7.1)$$

The existence of economies of output density ($E_{OD} > 1$) implies that the average cost of water distribution utility decreases as the physical output increases. Further, if the average cost decreases as the output and number of customers are proportionally increased, then

economies of customer density exist ($E_{CD} > 1$). Using Eq.(6.2), economies of customer density are calculated as:

$$\begin{aligned} E_{CD} &= \left(\frac{\partial \ln C}{\partial \ln Q} + \frac{\partial \ln C}{\partial \ln CU} \right)^{-1} \\ &= (b_Q + b_{Q,Q} \ln Q + b_{Q,CU} \ln CU + b_{Q,AS} \ln AS + d_{PL,Q} \ln PL^* + d_{PM,Q} \ln PM^* \\ &+ b_{CU} + b_{CU,CU} \ln CU + b_{Q,CU} \ln Q + b_{CU,AS} \ln AS + d_{PL,CU} \ln PL^* + d_{PM,CU} \ln PM^*)^{-1}. \end{aligned} \tag{7.2}$$

Finally, economies of size exist when $E_S > 1$ and are obtained from Eq. (6.2) in the following way:

$$\begin{aligned} E_S &= \left(\frac{\partial \ln C}{\partial \ln Q} + \frac{\partial \ln C}{\partial \ln CU} + \frac{\partial \ln C}{\partial \ln AS} \right)^{-1} \\ &= (b_Q + b_{Q,Q} \ln Q + b_{Q,CU} \ln CU + b_{Q,AS} \ln AS + d_{PL,Q} \ln PL^* + d_{PM,Q} \ln PM^* \\ &+ b_{CU} + b_{CU,CU} \ln CU + b_{Q,CU} \ln Q + b_{CU,AS} \ln AS + d_{PL,CU} \ln PL^* + d_{PM,CU} \ln PM^* \\ &+ b_{AS} + b_{AS,AS} \ln AS + b_{Q,AS} \ln Q + b_{CU,AS} \ln CU + d_{PL,AS} \ln PL^* + d_{PM,AS} \ln PM^*)^{-1}. \end{aligned} \tag{7.3}$$

As already noted in Section 3.5.2, economies of size and economies of scale do not necessarily correspond. Economies of size measure the percentage increase in cost due to a proportional increase in output, number of customers and size of the service area, while economies of scale measures a percentage increase in output as a result of proportional increase in all inputs. These two measures only correspond in the case of a homothetic production function (Chambers, 1988). However, in the applied literature (also recall the literature review in Section 6.1) the authors do not make a distinction between economies of size and economies of scale. They speak of economies of scale but they are in fact estimating economies of size. Thus, in what follows we will not maintain a strict distinction between the two expressions.

Estimated economies of output density, customer density and economies of scale for Slovenian water distribution utilities can be found in Table 7.5. The respective measures for all six models are calculated using Eq.(7.1), Eq.(7.2) and Eq.(7.3), where the input prices are held fixed at their median values. With respect to the amount of water distributed, the number of customers, and the size of service area three types of representative companies are chosen – a first-quartile company (small companies), a median company (medium-sized companies) and a third-quartile company (large companies). Based on the discussion in Section 7.1, the TFE model is believed to be our reference model. Here the results from

different models as reported in Table 7.5 demonstrate far more consistency than in the case of cost-inefficiency scores. All results follow the same pattern and, except for Model VI, lead us to the same conclusions. It can be noticed that the respective measures estimated by Model VI are considerably higher than those obtained by other models. This can again be attributed to the very low within variation in our model. As a result, the obtained estimates of Model VI may be imprecise.

Table 7.5: Economies of output density (E_{OD}), customer density (E_{CD}) and scale (E_S)

Economies	Quartile	*Model I* COLS	*Model II* Pooled	*Model III* RE (GLS)	*Model IV* RE (ML)	*Model V* TFE	*Model VI* RE (GLS) + Mundlak
E_{OD}	1st Quartile	2.839	3.099	3.485	3.500	4.605	5.041
	Median	2.767	3.042	3.455	3.448	3.874	6.380
	3rd Quartile	1.934	1.846	2.509	2.689	2.029	5.503
E_{CD}	1st Quartile	1.190	1.214	1.222	1.277	1.109	1.607
	Median	1.259	1.286	1.316	1.344	1.313	2.198
	3rd Quartile	1.172	1.182	1.265	1.263	1.208	2.809
E_S	1st Quartile	1.239	1.289	1.121	1.157	1.311	2.077
	Median	1.035	1.030	1.040	1.039	1.088	1.526
	3rd Quartile	0.854	0.816	0.933	0.925	0.846	1.138

Economies of output density (E_{OD}) are present for all three types of companies with respect to size. Since $E_{OD} > 1$, a 1% increase in cost (C) is associated with a more than 1% increase in the amount of water distributed (Q), holding the number of customers (CU) and the size of the service area (AS) constant. It would therefore be beneficial for water companies if they managed to distribute larger amounts of output to the existing customers within their service areas. E_{OD} are the highest for small utilities, followed by medium-sized utilities and large utilities. In Model VI, E_{OD} are highest for the median company. The economies of customer density (E_{CD}) are also confirmed for all three different types of companies. A 1% proportional increase in both the output and the number of customers leads to an increase in cost by less than 1% ($E_{CD} > 1$), holding the area size constant. Thus, it would be beneficial for companies if the existing service area were to become more densely populated or if the companies could manage to get new customers. E_{CD} are the highest for the median company (in Model VI they are the highest for the third-quartile company).

The economies of scale (E_S) equal the inverse of the percentage change in costs when the output, number of customers and area size increase by 1%. The results show that substantial

economies of scale are present in smaller companies ($E_S > 1$). It would be thus rational for the smaller companies to merge. Economies of scale are also present in medium-sized companies, where they are close to one. This is also an indication that the optimal size of Slovenian water distribution utilities is relatively close to the median point of the sample. The median company corresponds to a company with an annual water supply of 1,17 million cubic metres, 5,168 customers and 264 square kilometres of service area size. On the other hand, diseconomies of scale prevail in large companies ($E_s < 1$). Only Model VI finds economies of scale in all three cases. Apparently, the largest water distribution utilities in the sample have already exhausted their potential for cost savings resulting from economies of scale and their operations are found to be on the interval where average costs already start to rise.

7.4 Total Factor Productivity Growth Decomposition

Another way in which the cost frontier function may be employed is to decompose total factor productivity growth. Measures of productivity and associated productivity growth are also of great interest when analysing firm performance. Total factor productivity growth is one of the most widely employed measures of overall productivity change. Productivity growth also plays an important role in incentive-based price regulation. Recall the discussion of price-cap (RPI-X) regulation in Section 2.2.2 where it was established that the X factor is typically set to reflect the expected growth in total factor productivity based on past TFP growth. Therefore, regulatory authorities may be interested in measuring TFP growth and in determining those components that make the most significant contribution to the TFP growth.

Following the index number approach, a TFP index is generally constructed as the ratio of an output index to an input index where the weights reflect the relative importance of the various inputs and outputs (i.e., the weights equal the revenue shares and cost shares, respectively). In a single output case, TFP growth in defined as (Jorgenson and Griliches, 1967):

$$T\dot{F}P = \dot{y} - \dot{F} = \dot{y} - \sum_{i=1}^{K} \frac{w_i x_i}{C} \dot{x}_i , \tag{7.4}$$

where a dot over a variable indicates its rate of growth: $\dot{z} = d \ln z / dt = (1/z)(dz/dt)$. The observed output is denoted by y, w_i is the i-th input price, x_i is the observed use of i-th input,

C is the observed cost, and F stands for an aggregate measure of an observed input usage, with weights equalling the observed cost shares of the inputs used.

In order to decompose TFP growth, we apply a cost function approach.[117] Here the cost frontier function is, in addition, allowed to be a function of time so $C^* = C\ (y,\ \mathbf{w},\ t)$. Accordingly, the definition of cost efficiency in Eq.(3.24) can be rewritten as:

$$CE = C(y, \mathbf{w}, t) / C\ . \tag{7.5}$$

Taking the natural logarithm of both sides of Eq.(7.5) and totally differentiating with respect to time yields:

$$\dot{CE} = \varepsilon_y(y, \mathbf{w}, t)\ \dot{y} + \sum_{i=1}^{K} \frac{\partial C(y, \mathbf{w}, t)}{\partial w_i} \frac{w_i}{C(y, \mathbf{w}, t)} \dot{w}_i + \dot{C}(y, \mathbf{w}, t) - \dot{C}\ , \tag{7.6}$$

where $\varepsilon_y\ (y,\ \mathbf{w},\ t)$ is the elasticity of cost with respect to the output, as defined in Eq.(3.33). Using the definition of TFP growth in Eq.(7.4) and making some minor substitutions and rearrangements of Eq.(7.6), the decomposition of the observed TFP growth can be written as follows (Bauer, 1990):

$$\dot{TFP} = \dot{CE} - \dot{C}(y, \mathbf{w}, t) + \left[1 - \varepsilon_y(y, \mathbf{w}, t)\right] \dot{y} + \sum_{i=1}^{K} \left[S_i - S_i(y, \mathbf{w}, t)\right] \dot{w}_i\ , \tag{7.7}$$

where $S_i\ (y,\ \mathbf{w},\ t) = w_i\ x_i(y,\ \mathbf{w},\ t)\ /\ C(y,\ \mathbf{w},\ t)$ is the cost-minimising cost share of i-th input, as defined in Eq.(3.8), and $S_i = w_i\, x_i / C_i$ is the observed cost share of the i-th input. According to Eq.(7.7) TFP growth is decomposed into terms related to: (i) cost efficiency change; (ii) technical change; (iii) scale efficiency change; and (iv) a residual price effect term.

The first component captures the contribution to productivity change of change in cost efficiency, which is composed of a technical and allocative efficiency change. The second

[117] There are two main ways to derive TFP growth decomposition, the total differential method (see, for example, Bauer, 1990, and Kumbhakar and Lovell, 2000) and the index number method (see, for example, Caves, Christensen and Diewert, 1982, and Orea, 2002). The two approaches result in almost identical formulas, the only difference being that the first approach chooses just one data point in time for derivative evaluation, while the latter approach evaluates derivatives at two data points. Here, consistent with the previous analysis a differential approach is used.

component is a technical change effect that shifts the cost frontier down if technological progress is present, or up if technical change is regress. The third component is a scale effect which makes no contribution to productivity change if either the elasticity of cost with respect to the output equals one or there is no change in the output produced. Output growth in the presence of scale economies (ε_y (y, $\mathbf{w}$, t) < 1) contributes to productivity growth, as does output contraction in the presence of diseconomies of scale (ε_y (y, $\mathbf{w}$, t) < 1). Conversely, output growth in the presence of diseconomies of scale retards productivity growth, as does output contraction in the presence of economies of scale (Kumbhakar and Lovell, 2000).

The fourth component, the price effect term, occurs because the aggregate measure of input usage is biased when the firm is allocatively inefficient. If the firm is allocatively efficient, then $S_i = S_i$ (y, $\mathbf{w}$, t) and the price effect term is equal to zero.[118] The price term effect is present because TPF is defined as an observable quantity and therefore relies on observed input usage which might be biased due to the cost inefficiency. Alternatively, an unbiased or pure measure of TFP growth could be defined by omitting the price effect term but the link to an observable quantity would be lost (Bauer, 1990).

The TFP growth decomposition can be extended to the multiple output case. For the multi-product firm, the rate of growth of TFP can be defined in the following way (Jorgenson and Griliches, 1967):

$$T\dot{F}P = \dot{y}^P - \dot{F} = \sum_{j=1}^{M} \frac{p_j y_j}{R} \dot{y}_j - \sum_{i=1}^{K} \frac{w_i x_i}{C} \dot{x}_i , \tag{7.8}$$

where y^P is a revenue-weighted index of output, $\mathbf{y}$ is now a vector of outputs, $\mathbf{p}$ is a vector of output prices, $R = \mathbf{p}^{\mathbf{T}} \mathbf{y}$ is total revenue, and everything else is defined as before.

Using the same basic steps and manipulations as in the single-output case, the observed TFP growth for a multi-product firm can be shown to equal the following (Bauer, 1990):

[118] This term is also equal to zero when input prices change at the same rate, since $\sum_i [S_i - S_i(y, \mathbf{w}, t)] = 0$.

$$\dot{TFP} = \dot{CE} - \dot{C}(y,\mathbf{w},t) + \left[1 - \sum_{j=1}^{J} \varepsilon_{y_j}(y,\mathbf{w},t)\right]\dot{y}^C + \sum_{i=1}^{K}\left[S_i - S_i(y,\mathbf{w},t)\right]\dot{w}_i + (\dot{y}^P - \dot{y}^C), \tag{7.9}$$

where $\dot{y}^C = \sum_j \varepsilon_{y_j}\dot{y}_j \Big/ \sum_j \varepsilon_{y_j}$ and $\varepsilon_{y_j} = \partial \ln C(\mathbf{y},\mathbf{w},t)/\partial \ln y_j$.

In a similar manner, Eq.(7.9) decomposes TFP growth into change in cost efficiency, technical change, scale efficiency change, the price effect term, and an additional term capturing the effect that non-marginal cost pricing may have on the observed measure of TFP. The last effect occurs when observed revenue shares are not equal to the output elasticity shares (Bauer, 1990). The TFP decomposition thus provides useful conceptual and empirical tools for assigning the observed changes in TFP to various sources.

For network industries, output characteristics have an important influence on the cost of providing a certain output. Thus, these characteristics are incorporated in the cost function and also have to be taken into the account in the TFP growth decomposition. To allow for the effect of output characteristics on the TFP growth, Eq.(7.7) and Eq.(7.9) have to be properly modified (Bauer, 1990). In our single-output case with two output characteristics, the TFP growth decomposition is obtained in the following way:

$$\dot{TFP}_{it} = \dot{CE}_{it} - \frac{\partial \ln \hat{C}_{it}}{\partial t} + (1-\varepsilon_Q)\dot{Q}_{it} - \varepsilon_{CU}\,\dot{CU}_{it} - \varepsilon_{AS}\,\dot{AS}_{it} + (LS_{it} - LS_{it}^*)\dot{L} + (MS_{it} - MS_{it}^*)\dot{M} + (KS_{it} - KS_{it}^*)\dot{K}. \tag{7.10}$$

The first term on the left-hand side of Eq.(7.10) represents the cost efficiency change (CEC), the second term embodies the technical change (TC), the third term characterises the scale efficiency change (SEC), the fourth and fifth terms correspond to a change in output characteristics (OCC), while the last three terms capture the residual price effect (PER). If an increase in a given network characteristic increases (decreases) the cost given that the output remains unchanged, then increasing the level of that variable decreases (increases) TFP growth.

$\hat{C}_{it}$ is the predicted total cost obtained by estimating Eq.(6.2). Here, time variable *t* is considered to be a neutral technical change. Interactions of *t* with other variables are not

considered since insignificant coefficients were obtained.[119] In estimating the translog cost frontier function, Model V (TFE model) is employed since its performance compared to the other models is found to be superior with respect to both estimated coefficients and inefficiency scores. All components of the TFP growth can be then obtained from the estimated cost frontier function. *CE* is a cost-efficiency score which is the inverse of cost inefficiency and is obtained from the estimated cost frontier function by using Eq.(5.3). Technical change is calculated by taking the derivative of estimated cost frontier function with respect to time and in our case equals to h_t. Further, ε_Q, ε_{CU} and ε_{AS} are elasticities of cost with respect to the output delivered (*Q*), number of customers (*CU*) and area size (*AS*), respectively. *LS*, *MS* and *KS* stand for the observed cost shares of labour (*L*), material (*M*) and capital (*K*), while LS^*, MS^* and KS^* are the respective cost-minimising shares obtained by taking the derivative of the estimated cost frontier with respect to the price of labour, material and capital.

The decomposition of the TFP growth of Slovenian water distribution utilities in the 1997-2003 period based on the TFE model is reported in Table 7.6. Since the mean values can be influenced by the extreme values, mean and median yearly percentage growth rates are reported. Outliers can be noticed from the histograms of TFP growth which can be found in Appendix VI (Figures VI.1-VI.6). Such histograms are provided for all six models. From the histograms it can also be seen that the total factor productivity growth for the sample of Slovenian water distribution utilities is concentrated around a zero rate of growth.

From Table 7.6 it can be observed that cost efficiency in the TFE model practically remained unchanged. The mean cost efficiency change does not significantly differ from zero, implying that this component did not contribute to the TFP growth. Further, the technical change had a positive contribution to the TFP growth as costs are found to have been decreasing over the examined period. Annual technical progress of 0.77% is established, with this effect being significantly different from zero under the one-sided test. With respect to scale efficiency change, a positive contribution to the TFP growth is found. In absolute terms, this effect does not have a notable contribution to the TFP growth and is not shown to be significantly different from zero. On the other hand, changes in output characteristics have a stronger influence on the TFP growth. A change in output characteristics is found to be positive and significant, with this resulting in higher total costs and therefore a negative contribution to the TFP growth. Apparently, the companies have expanded their networks to include less populated areas, this resulting in less than a proportional increase of output supplied. With respect to the residual price effect, the median percentage growth rate is

[119] For example, Bauer (1990) also considered a neutral technical change.

found to be negative while the mean value is estimated to be positive. This is due to the fact that the median as opposed to the mean is not affected by the outliers. Nevertheless, the residual price effect is not found to be significant.

Table 7.6: TFP growth decomposition (average and median annual relative changes in percent)

TFP growth component	Model V (TFE)	
	Mean	Median
Cost efficiency change (*CEC*)	0.077	-0.090
Technical change (*TC*)	-0.771*	-0.771
Scale efficiency change (*SEC*)	0.187	0.143
Change in output characteristics (*OCC*)	0.794**	0.292
Residual price effect (*PER*)	0.497	-0.255
TFP growth (*TFPC*)	0.739	0.277
Pure TFP growth (without *PER*)	0.241	0.532

Note: * – significant at 10% (two-sided significance level);
** – significant at 5% (two-sided significance level)

Putting all these effects together, we obtain TFP growth which is found to be slightly increasing over the examined period. The same conclusion can be made if we observe only pure TFP growth, i.e. without the residual price effects. However, the established positive TFP growth is not found to be significant. It seems that the two components that have significantly contributed to the TFP growth over the examined period, namely technical progress and changes in output characteristics, cancel each other out since the first has a positive contribution while the contribution of the second component is negative. On the other hand, cost efficiency improvements are not found to have contributed to the TFP growth. Once again, this is in line with the absence of proper incentives to stimulate Slovenian water distribution utilities to operate in a more efficient way.

8 Conclusions

Despite the ongoing reforms of network industries in the direction of market liberalisation and privatisation, some form of regulation is still needed in order to prevent the abuse of a monopoly position and to enhance the efficient production of natural monopolies operating within network industries. In this respect, incentive-based regulation schemes appear to be superior to the traditional rate-of-return regulation. While the incentive-based price-cap (RPI-X) mechanism provides firms with incentives for efficient production it is not, however, without its shortcomings. If firms recognise that prices ultimately follow costs, they may well not reduce costs to efficient levels. To remove or at least lessen the informational asymmetry problem between regulators and firms operating in regulated industries, the regulator needs reliable estimates of the efficiency potential of a regulated firm. This can be obtained by performing some form of benchmarking analysis.

In the thesis we examined several parametric frontier benchmarking methods to estimate firms' (in)efficiency. The main methodology chosen to be applied was Stochastic Frontier Analysis (SFA), where a special focus is put on the panel data stochastic frontier models. The models are found to differ in distributional assumptions required and in the method of estimation employed. The models also differ in their ability to account for firm-specific effects and to distinguish between firm heterogeneity and inefficiency. While conventional fixed and random effects panel data models take firm-specific effects into account in the estimation of inefficiency, they are unable to distinguish between unobserved heterogeneity and inefficiency. Any time-invariant firm-specific effects are treated as inefficiency. Although this can have a huge influence on the estimated cost inefficiencies, this problem has been ignored for a long time. In network industries, the problem of unobserved heterogeneity is even more severe since utilities operate in different regions that typically differ in their environmental and network characteristics. It has only been recently recognised that it is important to be able to separate unobserved heterogeneity from inefficiency. An interesting extension of the conventional panel data models is Mundlak's (1978) formulation of a random effects model which can be employed to control for the correlation between unobserved heterogeneity and regressors. The newly proposed 'true' fixed and random effects models proposed by Greene (2002a, b) attempt to distinguish between unobserved heterogeneity and inefficiency by adding an additional term into the model which is supposed to capture time-invariant and firm-specific effects. Nevertheless, this novelty does not fully resolve the problem since any time-invariant firm-specific effects are treated as unobserved heterogeneity, including time-invariant inefficiency. As the

conventional FE and RE models tend to overestimate the inefficiency, it may be the case that the TFE and TRE models underestimate it.

In the empirical part of the thesis, our aim was to estimate the cost frontier function of Slovenian water distribution utilities in order to estimate their cost inefficiency and, further, to consider the possible use of the results in the price regulation of these utilities. The Slovenian water industry is currently facing a range of problems. At present, the prices for delivering water in Slovenia vary significantly between local communities and typically do not reach the full-cost level. The current price regulation resembles the rate-of-return regulation. In addition, the upper limit on permissible price increases is fixed by a special decree. In order to comply with the EU legislation relating to water policy and water pricing, the reform of the Slovenian water industry is inevitable. Based on the presented best-practice examples of water price regulation from Italy and the UK, the most important tasks to be accomplished in the Slovenian water industry appear to be the introduction of cost-reflective prices of supplied water, new investments in the distribution network, and establishing a new (incentive-based) regulatory framework. An independent regulatory authority should be established to supervise water distribution utilities, which is currently the task of the relevant ministries and can thus be subjected to policy considerations. In setting prices for the water supplied, an incentive-based scheme should be introduced to provide utilities with incentives to reduce their costs and increase their efficiency. In fact, the currently issued Rules on Price Determination of Obligatory Local Public Utilities for Environment Protection (2004) envisage the use of determination of prices based on justified costs, the identification of the best-practice performance and benchmarking, but they have not yet been put into practice.

In the thesis, the possible use of parametric frontier benchmarking methods for price-regulation purposes is explored. Several stochastic frontier methods were employed to estimate the cost inefficiency of Slovenian water distribution utilities. A translog total cost frontier function was employed on an unbalanced panel data set of 52 utilities over the 1997-2003 period. The estimation results suggest that considerable cost inefficiency is present in Slovenia's water distribution companies. In estimating inefficiency, our main objective was to take the unobserved heterogeneity into account and to analyse how the results are influenced by the separation of unobserved heterogeneity from inefficiency. Conventional RE models are found to highly overestimate cost inefficiency since the inefficiency estimates also contain unobserved heterogeneity. The TFE model is able to distinguish between unobserved heterogeneity and inefficiency but it may slightly underestimate the inefficiency since all time-invariant effects are treated as unobserved heterogeneity. Nevertheless, since the inefficiency estimates obtained by the TFE model closely correspond to the pooled model it is believed that the mean cost inefficiency of Slovenian water distribution utilities is

close to or slightly above 20%. Also, by taking into consideration the expected signs and significance of the obtained coefficients of the cost frontier function the TFE model is found to perform better than the other models and is therefore chosen as our preferable model.

Further, the inefficiency scores obtained from different methods are not found to be consistent in their levels and rankings of the utilities. The established inconsistency of the inefficiency estimates from different models is not specific to our sample but is quite common in the applied economic literature. A possible explanation of the inconsistent results can be found in the differences seen in stochastic frontier methods when accounting for unobservable heterogeneity. Nevertheless, if SFA is supposed to serve as a benchmarking tool in price regulation the inconsistency of the different models is particularly undesirable. If different methods produce very different results for inefficiency scores, we cannot rely on the findings from SFA and use them directly in economic policy-making. In this case, SFA as a benchmarking tool should only be used as a complementary instrument for setting efficiency requirements as part of price regulation. However, benchmarking can still be viewed as a useful instrument for reasonably mitigating the informational asymmetries between the regulator and utilities. Since the regulators can only imperfectly observe the utilities' performance, the benchmarking results should merely be viewed as a starting point in providing information about the range in which the inefficiency can be located. This is largely consistent with the practice of the UK regulator OFWAT.

Since the cost function estimated by the GLS method is found to predict the total cost of companies with reasonable precision, the results can be alternatively used by the regulator to predict utilities' costs and to thus set the interval for allowable costs. This solution essentially implies reverting from frontier benchmarking to yardstick competition as proposed by Shleifer (1985) and is in line with the practice followed by the Italian Regulation Authority. When applying this approach to the price-regulation process, the utilities may be given the possibility to renegotiate prices with the regulator by justifying higher than expected costs. This situation would be even more likely if the model is not shown to have a very strong predictive power.

Finally, with respect to economies of scale and density the results are more consistent. The estimated economies of scale close to one for the sample median point indicate that medium-sized utilities closely correspond to the optimal size of water distribution utilities in Slovenia. Large utilities are found to operate at levels where diseconomies of scale are already present, while smaller utilities should be interested in mergers since this would lead to a decrease in average operating costs. Economies of output density and customer density are confirmed for all three different types of utilities with respect to the size of the operation.

Therefore, it would be beneficial for the utilities if they managed to increase the volume of water supplied to their existing customers as well as to acquire new customers.

Hence, significant scope for cost reductions exists within the water distribution utilities. Two possible sources of cost savings in the Slovenian water industry are recognised, namely improving on cost efficiency and scale efficiency. So far, no significant improvements in this direction have been made. This is confirmed by the estimated total factor productivity (TFP) growth which is found to be concentrated around zero. TFP growth is also an important measure for analysing firm performance and it is widely used in price-cap regulation to set the productivity requirements of regulated companies. The estimated cost frontier function can be alternatively used to decompose TFP growth into different components and to establish their contribution to TFP growth. The results based on the TFE model suggest that the cost efficiency of Slovenian water distribution utilities did not improve significantly over the examined period, whereas technical progress did have a marked contribution on TFP growth. Overall, TFP growth in the water distribution utilities is not found to be significantly different from zero. Once again it is confirmed that the present non-competitive environment in which utilities operate as well as the current regulatory framework do not provide utilities with sufficient incentives for improving their efficiency and reducing their costs. In order to facilitate the improved performance of water distribution utilities a regulatory reform is urgently needed.

From the scientific point of view, the main findings of the thesis are the following:

- The inclusion of output characteristics in the cost function is important for modelling the production process of utilities operating in a network industry and it allows for a distinction between economies of density and economies of scale. The different models are found to be fairly consistent in estimating economies of output density, customer density and economies of scale.
- In the stochastic frontier analysis it is important to account for firm heterogeneity and to be able to distinguish between unobserved heterogeneity and inefficiency. If not, a heterogeneity bias may result in both biased inefficiency estimates as well as biased coefficient estimates. However, it remains debatable whether certain firm-specific effects should in fact be attributed to unobserved heterogeneity or inefficiency.
- The inefficiency scores obtained from the different cost frontier models are not found to be consistent. The levels of inefficiency estimates as well as the rankings depend on the econometric specification of the model. The established inconsistency can at least

to some extent be contributed to the different ability of the models to separate unobserved heterogeneity from inefficiency.

From an economic policy point of view, the findings imply that:

- Reforms of the Slovenian water industry in the direction of introducing cost-reflective prices, an incentive-based regulation mechanism and an independent regulatory authority are needed.
- Significant cost inefficiencies are present in Slovenian water distribution utilities. An incentive-based price regulation might help resolve this problem.
- The inconsistency of the inefficiency scores obtained suggests that benchmarking should only be used as a complementary method in the price-setting process. The mechanical use of SFA results is not recommended. The benchmarking results can only be viewed as a starting point for providing information on the range in which the inefficiency scores can be located.
- The presence of economies of output and customer density in Slovenian water distribution utilities is established. Economies of scale are found in small-sized utilities, median-sized utilities demonstrate economies of scale close to one, while large companies exhibit diseconomies of scale. The optimal size of a company thus closely corresponds to the sample median while, in order to exploit economies of scale, it would make sense to merge the smaller utilities.

Appendix I

Table I.1: Prices of water supplied in Slovenia in the period 1991-2000

Variable	31.6.91	31.12.91	31.12.92	31.12.93	31.12.94	31.12.95	31.12.96	30.4.97	31.12.98	31.12.99	31.12.00
Average household prices P_H (SIT/m3)[1,2]	10,2	15,5	28,1	35,3	45,4	51,8	55,9	60,1	64,0	74,8	87,7
Average business prices P_B (SIT/m3)	20,9	31,4	55,6	69,2	85,8	98,6	104,6	109,7	113,8	122,7	136,3
Ratio of max to min price for households	10,5	9,7	13,8	10,3	8,3	9,0	8,4	9,0	8,5	8,8	8,3
Ratio of max to min price for businesses	10,5	12,6	16,3	11,9	8,4	9,1	8,6	8,6	9,4	10,1	9,2
Ratio of household to business prices P_B/P_H	2,05	2,02	1,98	1,96	1,89	1,90	1,87	1,83	1,78	1,76	1,64
Inflation index (RPI)[3, 4]		223,9	192,9	122,9	118,3	108,6	108,8	103,8 (109,4)	107,5	108,8	110,6
Fixed household prices (30.4.1997=100)	78,7	53,4	50,0	51,2	55,7	58,5	60,8	60,1	56,5	50,3	54,0
Fixed business prices (30.4.1997=100)	161,0	107,9	99,2	100,4	105,2	111,3	113,8	109,7	100,4	88,7	88,6

[1] SIT stands for the Slovenian currency (Slovenian tolar). For international comparisons, the exchange rate at the end of 1991 was 1 EUR = 75,756 SIT, while at the end of 2000 the exchange rate rose to 1 EUR = 211,506 SIT (Bank of Slovenia, 1999, 2001).
[2] Prices of water supplied refer to the end prices paid by the customers and include other items as well (contribution in addition to the price, state fee, local fee and other items).
[3] Until 1998, official rate of inflation in Slovenia was measured by the Retail Price Index (RPI); from 1998, the official inflation is expressed by the Consumer price index (CPI).
[4] Inflation indices provided in the table above refer to the general price level at the end of the current year relative to the end of the previous year. There are two exceptions. For 1991, the inflation index refers to the inflation rate in the second half of the year, which amounted to 123,9%. For 1997, the inflation rate for the first four months was 3,8%, while the inflation rate at the end of 1997 compared to the end of 1996 was 9,4%.

Source: Štruc (1997), Svetovalni center (1997, 2001), Hrovatin (2002) and Statistical Yearbook of the RS (2001)

Table I.2: Water supplied from the public water supply systems, length of the water distribution network and number of connections in Slovenia

Year	1990	1995	2000	2002
Total water supplied (1000 m^3)	262144	259687	235826	183421
Households	86217	86475	87968	88470
Businesses	79834	56294	46175	37559
To other water supply systems	16304	9631	7277	
Supplied but uncharged water				7376
Leakage	79789	107287	94406	50016
Network length (km)	13630	13433	16164	16598
Number of connections	328579	353164	406302	415763

Source: Statistical Yearbook of the RS 2003

Table I.3: Some relevant indicators of Slovenian water supply sector in the period 1992-2002

Year	1992	1993	1994	1995	1996	1997
Nr. of utilities	46	46	47	48	50	52
Nr. of employees	4385	4235	4216	4269	4202	3924
% of fixed assets in total assets			84,1	88,4	89,1	91,0
Sales (in SIT million)	12762	17380	22244	25168	27976	28291
Net Overall Profit(+)/Loss(-) (in SIT million)	1	375	229	-335	-459	-765
Net Profit(+)/Loss(-) from Regular Activity (in SIT million)	53	438	-196	-566	-1418	-1596

Table I.3: Continuation

Year	1998	1999	2000	2001	2002
Nr. of utilities	52	52	54	56	56
Nr. of employees	3866	3854	3786	3833	3896
% of fixed assets in total assets	90,5	90,7	91,0	90,7	89,6
Sales (in SIT million)	33547	35389	36889	42607	47168
Net Overall Profit(+)/Loss(-) (in SIT million)	-469	-879	-1422	-402	82
Net Profit(+)/Loss(-) from Regular Activity (in SIT million)	-1540	-2541	-2804	-1875	-2557

Source: Svetovalni center (1997), PASEF (1994-2002)

Appendix II

Table II.1: List of Slovenian water distribution utilities in the sample, their legal form, region and number of local communities they provide with drinking water

Nr.	Company[1]	Region	Legal form[2]	Nr. of municipalities[3]
1	VO-KA Ljubljana	1 Osrednjeslovenska	PU	6 (+2)
2	Mariborski vodovod	7 Podravska	PU	11 (+2)
3	Komunala Kranj	3 Gorenjska	PU	5 (+1)
4	Rižanski vodovod Koper	12 Obalno-kraška	PU	3
5	VO-KA Celje	4 Savinjska	PU	4
6	KP Velenje	4 Savinjska	PU	3
7	Komunala Novo mesto	2 Dolenjska	PU	5 (+1)
8	JKP Prodnik Domžale	1 Osrednjeslovenska	PU	6
9	KP Ptuj	7 Podravska	PU	16 (+6)
10	Kraški vodovod Sežana	12 Obalno-kraška	PU	4 (+1)
11	JKP Žalec	4 Savinjska	PU	6
12	KJP Murska Sobota	8 Pomurska	PU	5
13	KSD Ajdovščina	10 Goriška	PU	2
14	OKP Rogaška Slatina	4 Savinjska	PU	6
15	JEKO-IN Jesenice	3 Gorenjska	PU	2
16	Hydrovod Kočevje	1 Osrednjeslovenska	PU	5
17	KSP Kostak Krško	6 Spodnjesavska	**P**	1 (+2)
18	JKP Grosuplje	1 Osrednjeslovenska	PU	2 (+1)
19	Loška Komunala	3 Gorenjska	**C**	1
20	Komunala Slovenska Bistrica	7 Podravska	PU	2 (+2)
21	Kovod Postojna	11 Notranjsko-kraška	PU	2
22	Komunala Radovljica	3 Gorenjska	PU	1
23	KSP Brežice	6 Spodnjesavska	PU	1
24	Komunala Tolmin	10 Goriška	PU	3
25	KP Vrhnika	1 Osrednjeslovenska	PU	2
26	KSP Hrastnik	5 Zasavska	PU	1
27	KP Ilirska Bistrica	11 Notranjsko-kraška	PU	1
28	JKP Log - Ravne	9 Koroška	PU	3
29	Komunala Trbovlje	5 Zasavska	PU	1
30	JKP Slovenj Gradec	9 Koroška	PU	2
31	Infrastruktura Bled	3 Gorenjska	PU	1
32	JPK Črnomelj	2 Dolenjska	PU	2
33	KP Ormož	7 Podravska	**C**	1 (+1)
34	KSP Ljutomer	8 Pomurska	PU	4 (+1)
35	JKP Slovenske Konjice	4 Savinjska	PU	1
36	Komunala Trebnje	2 Dolenjska	PU	1
37	Komunala Lendava	8 Pomurska	**C**	1 (+1)

38	KP Tržič	3	Gorenjska	PU	1 (+2)
39	JPK Cerknica	11	Notranjsko-kraška	PU	2
40	JKP Šentjur	4	Savinjska	PU	1
41	KP Logatec	1	Osrednjeslovenska	PU	1
42	KOP Zagorje	5	Zasavska	PU	1
43	KSP Litija	1	Osrednjeslovenska	PU	2
44	JPK Mozirje	4	Savinjska	PU	4
45	Komunala Idrija	10	Goriška	PU	1
46	Komunala Kranjska Gora	3	Gorenjska	PU	1
47	Komunala Metlika	2	Dolenjska	PU	1
48	JPK Sevnica	6	Spodnjesavska	PU	1
49	JKP Brezovica	1	Osrednjeslovenska	PU	1
50	JPK Radeče	4	Savinjska	PU	1 (+1)
51	JKP Dravograd	9	Koroška	PU	1
52	Komunala Gornji Grad	4	Savinjska	PU	1

[1] Companies are ranged from the biggest to the smallest one in terms of water supplied to the customers in 2003.

[2] Meaning of abbreviations: PU – public utility, P – private company, C – concession

[3] The number of municipalities that are only partially served by the respective company is given in brackets. The total number of municipalities in Slovenia in 2003 was 193.

Appendix III

Table III.1: Estimation results of the Cobb-Douglas (C-D), general translog (GT) and hedonic translog (HT) cost function

Coefficient	C-D (OLS)	GT (OLS)	Coefficient	HT(NLS)
$\ln a$	11,920***	11,834***	$\ln a$	11,881***
	(0,024)	(0,029)		(0,025)
c_{PL}	0,704***	0,586***	c_{PL}	0,652***
	(0,024)	(0,025)		(0,022)
c_{PM}	0,179***	0,188***	c_{PM}	0,146***
	(0,022)	(0,023)		(0,019)
b_Q	0,522***	0,363***	b_Y	0,498***
	(0,037)	(0,062)		(0,031)
b_{CU}	0,317***	0,432***	-	-
	(0,042)	(0,062)		
b_{AS}	0,170***	0,172***	-	-
	(0,024)	(0,024)		
$c_{PL,PL}$	-	-0,110**	$c_{PL,PL}$	-0,136***
		(0,051)		(0,042)
$c_{PM,PM}$	-	-0,084**	$c_{PM,PM}$	-0,105***
		(0,038)		(0,034)
$c_{PL,PM}$	-	0,159***	$c_{PL,PM}$	0,174***
		(0,041)		(0,042)
$b_{Q,Q}$	-	0,427***	$b_{Y,Y}$	0,015***
		(0,149)		(0,004)
$b_{CU,CU}$	-	0,015	-	-
		(0,234)		
$b_{AS,AS}$	-	0,122*	-	-
		(0,063)		
$b_{Q,CU}$	-	-0,304*	-	-
		(0,179)		
$b_{Q,AS}$	-	0,012	-	-
		(0,082)		
$b_{CU,AS}$	-	0,171**	-	-
		(0,082)		
$d_{PL,Q}$	-	0,095	$d_{PL,Y}$	0,007
		(0,081)		(0,015)
$d_{PL,CU}$	-	-0,106	-	-
		(0,078)		
$d_{PL,AS}$	-	-0,013	-	-
		(0,039)		
$d_{PM,Q}$	-	-0,076	$d_{PM,Y}$	0,001
		(0,060)		(0,015)
$d_{PM,CU}$	-	0,090	-	-
		(0,066)		
$d_{PM,AS}$	-	0,023	-	-
		(0,037)		
g_S	0,114***	0,176***	g_S	0,128***
	(0,037)	(0,033)		(0,033)
g_U	0,120***	0,057**	g_U	0,066***
	(0,032)	(0,028)		(0,027)
g_{TREAT}	0,251***	0,118***	g_{TREAT}	0,167***
	(0,043)	(0,037)		(0,037)
g_{LOSL}	-0,273***	-0,176***	g_{LOSL}	-0,244***
	(0,034)	(0,029)		(0,029)

			Coefficient	C-D output aggregator
-	-	-		
-	-	-	h_Q (imposed)	1
-	-	-	h_{CU}	0,677**** (0,110)
-	-	-	h_{AS}	0,280**** (0,044)
logL	4,308	101,657	*logL*	66,194

Notes: standard errors in brackets; * – significant at 10%, ** – significant at 5%, *** – significant at 1%, **** – significant at 0.1% (two-sided significance level)

Table III.2: Results of the likelihood ratio test

LR test	Translog vs. C-D (OLS)	Translog vs. hedonic
$-2(\log L_R - \log L_U)$	194,70	70,93
Significance level	0,000	0,000

Appendix IV

Table IV.1: Estimation results of the true random effects model (TRE)*

Coefficient	*TRE model*	Coefficient	*TRE model*
$\ln a$	11,825****	σ_ω	0,2361****
	(0,009)		(0,0038)
c_{PL}	0,391****	σ_u	0,0006
	(0,007)		
c_{PM}	0,350****	σ_v	0,0684****
	(0,006)		
b_Q	0,258****	$\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$	0,0684****
	(0,016)		(0,0013)
b_{CU}	0,470****	$\lambda = \sigma_u / \sigma_v$	0,0092
	(0,016)		(0,0567)
b_{AS}	0,229****		
	(0,006)		
$c_{PL,PL}$	0,055****		
	(0,013)		
$c_{PM,PM}$	0,033****		
	(0,010)		
$c_{PL,PM}$	-0,002		
	(0,011)		
$b_{Q,Q}$	0,328****		
	(0,038)		
$b_{CU,CU}$	0,012		
	(0,060)		
$b_{AS,AS}$	0,149****		
	(0,017)		
$b_{Q,CU}$	-0,173****		
	(0,046)		
$b_{Q,AS}$	-0,034*		
	(0,020)		
$b_{CU,AS}$	0,022		
	(0,021)		
$d_{PL,Q}$	0,126****		
	(0,019)		
$d_{PL,CU}$	-0,152****		
	(0, 019)		
$d_{PL,AS}$	0,034****		
	(0,010)		
$d_{PM,Q}$	-0,107****		
	(0,014)		
$d_{PM,CU}$	0,128****		
	(0,015)		
$d_{PM,AS}$	-0,067****		
	(0,037)		
h_T	-0,002*		
	(0,001)		
g_S	0,044****		
	(0,009)		
g_U	0,077****		
	(0,007)		
g_{TREAT}	0,285****		
	(0,010)		
g_{LOSL}	-0,027****		
	(0,007)		

* The estimation method did not converge.

Appendix V

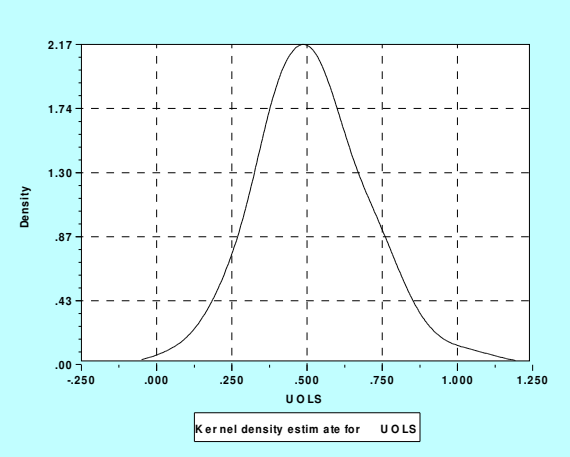

Figure V.1: Kernel density for Model I (COLS)

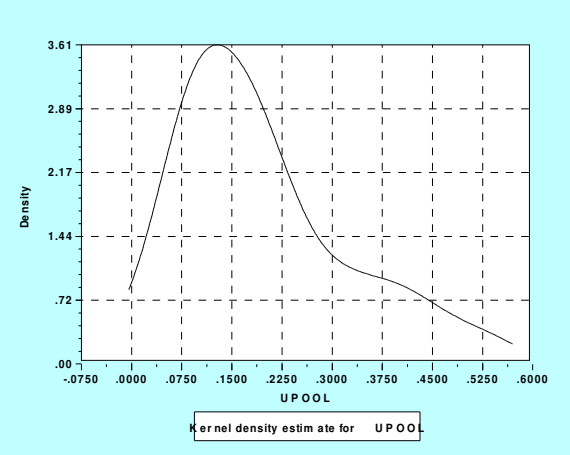

Figure V.2: Kernel density for Model II (Pooled MLE)

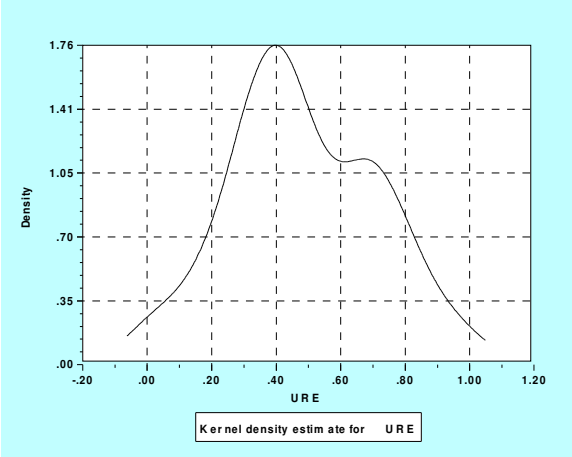

Figure V.3: Kernel density for Model III (RE GLS)

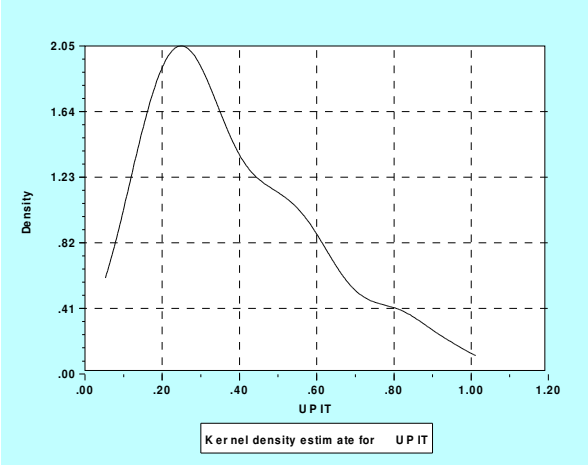

Figure V.4: Kernel density for Model IV (RE MLE)

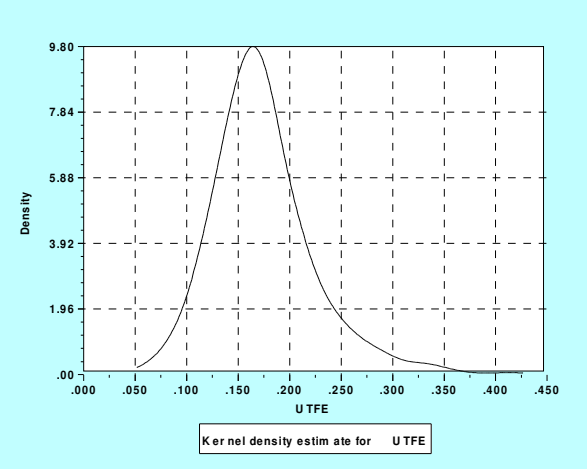

Figure V.5: Kernel density for Model V (TFE)

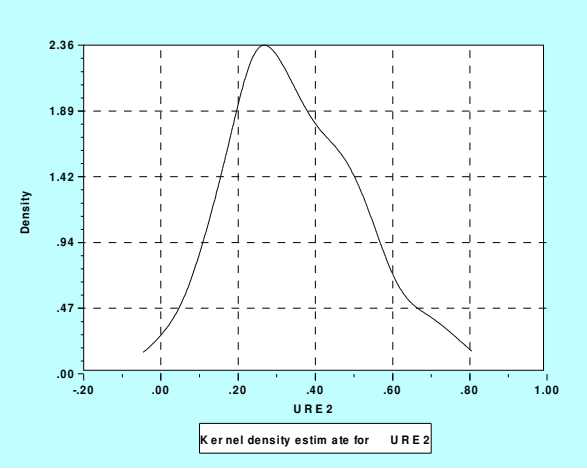

Figure V.6: Kernel density for Model VI (RE GLS + Mundlak)

Appendix VI

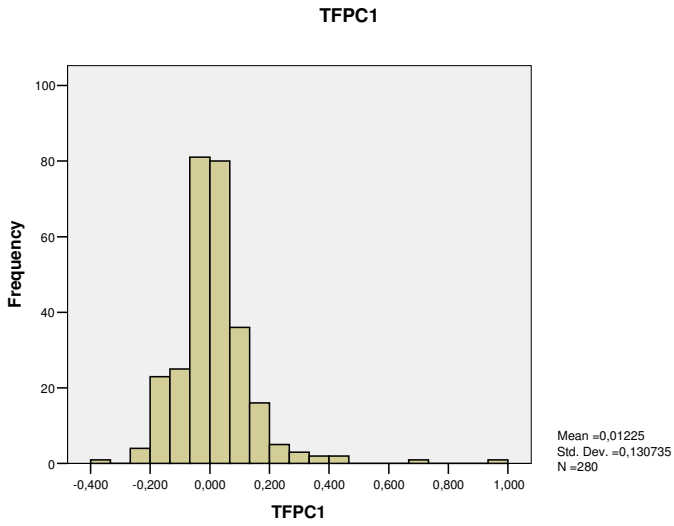

Figure VI.1: Histogram of TFP growth as calculated from Model I (COLS)

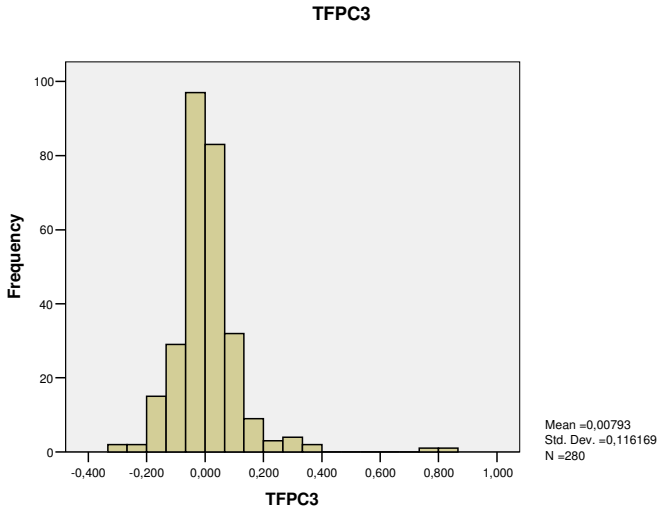

Figure VI.2: Histogram of TFP growth as calculated from Model II (Pooled MLE)

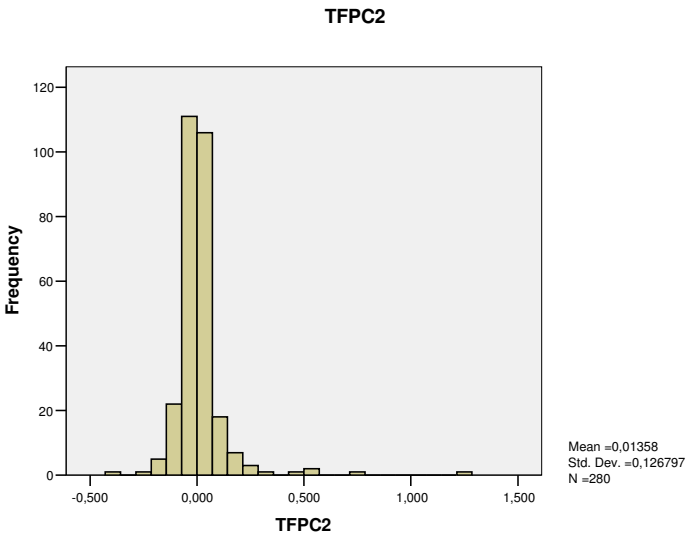

Figure VI.3: Histogram of TFP growth as calculated from Model III (RE GLS)

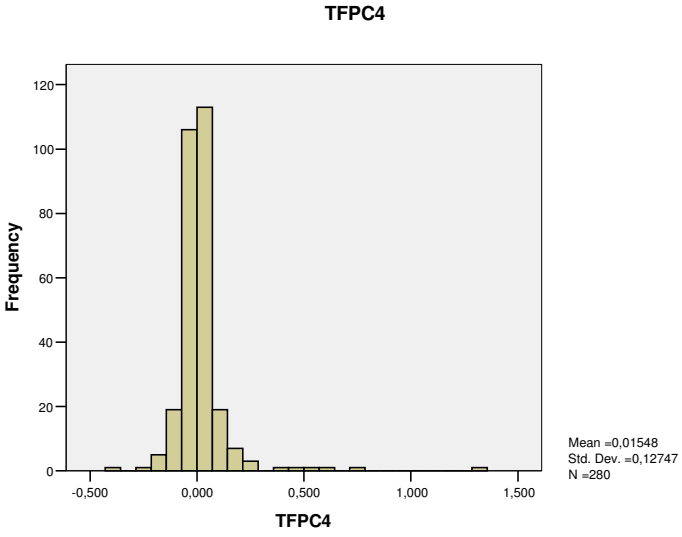

Figure VI.4: Histogram of TFP growth as calculated from Model IV (RE MLE)

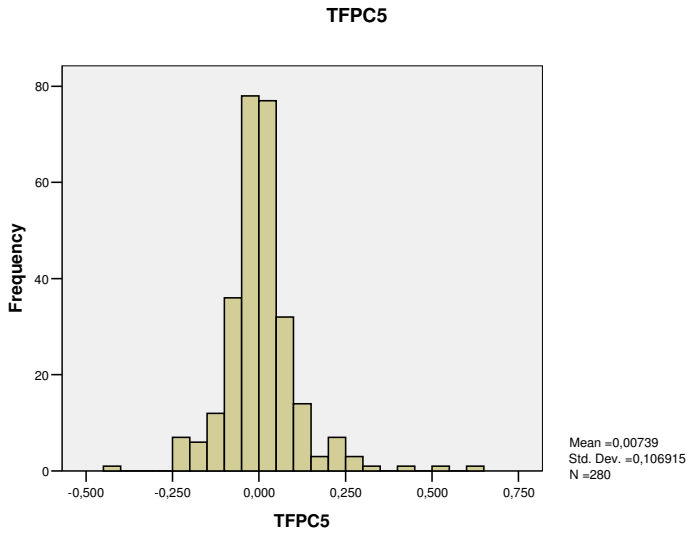

Figure VI.5: Histogram of TFP growth as calculated from Model V (TFE)

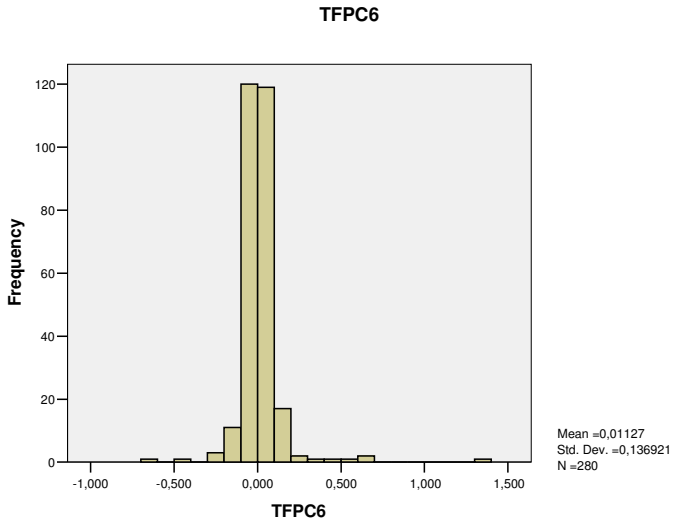

Figure VI.6: Histogram of TFP growth as calculated from Model VI (RE GLS + Mundlak)

Bibliography

Ahn, S.C., Lee, Y.H. and Schmidt, P. (1994). GMM Estimation of a Panel Data Regression Model with Time-Varying Individual Effects. Working Paper, Department of Economics, Michigan State University, East Lansing, MI.

Afriat, S.N. (1972). Efficiency Estimation of Production Functions. ***International Economic Review 13(3)***, 568-598.

Aigner, D.J., Lovell, C.A.K. and Schmidt, P. (1977). Formulation and Estimation of Stochastic Frontier Production Function Models. ***Journal of Econometrics 6(1)***, 21-37.

Allen, R.G.D. (1938). *Mathematical Analysis for Economists*. Macmillan, London.

Antonioli, B. and Filippini, M. (2001). The use of a variable cost function in the regulation of the Italian water industry. ***Utilities Policy 10***, 181-187.

Armstrong, M., Cowan, S., Vickers, J. (1994). *Regulatory Reform: Economic analysis and British experience*. MIT Press, Cambridge, Mass. and London.

Ashton, J.K. (2000). Cost efficiency in the UK water and sewerage industry. ***Applied Economics Letters 7***, 455-458.

Averch, H. and Johnson, L.L. (1962). Behavior of the Firm Under Regulatory Constraint. ***American Economic Review 52***, 1052-1069.

Baldwin, R., Cave, M. (1999). *Understanding Regulation: Theory, Strategy and Practice*. Oxford University Press, New York.

Bank of Slovenia (1999). Monthly Bulletin, Vol. 10, No. 1. January 1999, Ljubljana.

Bank of Slovenia (2001). Monthly Bulletin, Vol. 8, No. 1. January 2001, Ljubljana.

Banker, R.D., Charnes, A. and Cooper, W.W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. ***Management Science 30(9)***, 1078-1092.

Battese, G.E. and Coelli, T.J. (1992). Frontier Production Functions, Technical Efficiency and Panel Data: With Application to Paddy Farmers in India. ***Journal of Productivity Analysis 3(1/2)***, 153-169.

Battese, G.E. and Coelli, T.J. (1995). A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data. ***Empirical Economics 20***, 325-332.

Bauer, P.W. (1990). Decomposing TFP Growth in the Presence of Cost Inefficiency, Nonconstant Returns to Scale, and Technological Progress. ***The Journal of Productivity Analysis 1***, 287-299.

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

Baumol, W.J. (1977). On the Proper Cost Tests for Natural Monopoly in a Multi-Product Industry. ***American Economic Review 67(5)***, 809-822.

Baumol, W. (1982). Contestable Markets: An Uprising in the Theory of Industrial Structure. ***American Economic Review 72***, 1-15.

Bhattacharyya, A., Harris, T.R., Narayanan, R., Raffiee, K. (1995). Specification and estimation of the effect of ownership on the economic efficiency of the water utilities. ***Regional Science and Urban Economics 25***, 759-784.

Blackorby, C., Primont, D. and Russell, R.R. (1977). On Testing Separability Restrictions with Flexible Functioanl Forms. ***Journal of Econometrics 5***, 195-209.

Blackorby, C. and Russel, R.R. (1989). Will the Real Elasticity of Substitution Please Stand Up? A Comparison of the Allen/Uzawa and Morishima Elasticities. ***The American Economic Review 79(4)***, 882-888.

Box, G.E.P. and Cox, D.R. (1964). An Analysis of Transformations. ***Journal of the Royal Statistical Society, Series B, 26(2)***, 211-252.

Breyer, F. (1987). The specification of a hospital cost function: a comment on the recent literature. ***Journal of Health Economics 6(2)***, 147-157.

Bruggink, T.H. (1982). Public versus Regulated Private Enterprise in the Municipal Water Industry: A Comparison of Operating Costs. ***Quarterly Review of Economics and Business 22***, 111-125.

Cabral, L.M.B. and Riordan, M.H. (1989). Incentives for Cost Reduction under Price-Cap Regulation. ***Journal of Regulatory Economics 1(2)***, 93-102.

Cameron, A.C. and Trivedi, P.K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Caudill, S.B., Ford, J.M. and Gropper, D.M. (1995). Frontier Estimation and Firm-Specific Inefficiency Measures in the Presence of Heteroskedasticity. ***Economic Letters 41***, 17-20.

Caves, D.W., Christensen, L.R., Diewert, W.E. (1982). Multilateral Comparisons of Output, Input, and Productivity Using Superlative Index Numbers. ***Economic Journal 92(365)***, 73-86.

Caves, D.W., Christensen, L.R., Swanson, J.A. (1981). Productivity Growth, Scale Economies, and Capacity Utilization in U.S. Railroads, 1955-74. ***American Economic Review 71(5)***, 994-1002.

Caves, D.W., Christensen, L.R., Tretheway, M.W. (1980). Flexible Cost Functions for Multiproduct Firms. ***The Review of Economics and Statistics 62(3)***, 477-481.

Caves, D.W., Christensen, L.R., Tretheway, J.A. (1984). Economies of Density versus Economies of Scale: Why Trunk and Local Service Airline Costs Differ. ***Rand Journal of Economics 15(4)***, 471-489.

Chambers, R.G. (1988). *Applied Production Analysis*. Cambridge University Press.

Charnes, A., Cooper, W.W. and Rhodes, E. (1978). Measuring the efficiency of decision-making units. ***European Journal of Operational Research 2***, 429-444.

Christensen, L.R., Jorgenson, W.D., Lau, L.J. (1971). Conjugate Duality and the Transcendental Logarithmic Function. ***Econometrica 39***, 255-256.

Christensen, L.R., Jorgenson, W.D., Lau, L.J. (1973). Transcendental Logarithmic Production Frontiers. ***The Review of Economics and Statistics 55***, 28-45.

Clemenz, G. (1991). Optimal Price-Cap Regulation. ***Journal of Industrial Economics 39(4)***, 391-408.

Cobb, C.W. and Douglas, P.C. (1928). A Theory of Production. ***American Economic Review 18***, 139-165.

Coelli, T.J. (1995). Estimators and Hypothesis Tests for a Stochastic Frontier Function: A Monte Carlo Analysis. ***Journal of Productivity Analysis 6(4)***, 247-268.

Coelli, T., Estache, A., Perelman, S., Trujillo, L. (2003). A Primer on Efficiency Measurement for Utilities and Transport Regulators. WBI Development Studies, World Bank, Washington, D.C.

Coelli, T.J., Rao, P.D.S., Battese, G.E. (1998). *An Introduction to Efficiency and Productivity Analysis*. Kluwer Academic Publishers.

Communication from the Commission to the Council, the European Parliament and the Economic and Social Committee (2000). Pricing policies for enhancing the sustainability of water resources. COM(2000) 477 final. Commission of the European Communities, Brussels.

Cooper, W.W., Seinford, M.L., Tone, K. (2003). *Data Envelopment Analysis*. Kluwer Academic Publishers.

Cornes, R. (1992). *Duality and Modern Economics*. Cambridge University Press.

Cornvell, C., Schmidt, P. and Sickles, R.C. (1990). Production Frontiers with Cross-Sectional and Time-Series Variation in Efficiency Levels. ***Journal of Econometrics 46(1/2)***, 185-200.

Council Directive 98/83/EC of 3 November 1998 on the quality of water intended for human consumption. Official Journal of the European Communities L 330.

Cowing, T. and Holtmann A.G. (1983). Multiproduct Short-Run Hospital Cost Functions: Empirical Evidence and Policy Implications from Cross-Section Data. ***Southern Economic Journal 49(3)***, 637-653.

CPB Netherlands Bureau for Economic Policy Analysis (2000). Yardstick competition: Theory, design, and practice. Working Paper No 133, The Hague.

Crain, W.M. and Zardkoohi, A. (1978). A Test of the Property Rights Theory of the Firm: Water Utilities in the United States. ***Journal of Law and Economics 21***, 395-408.

Crew, M.A., Kleindorfer, P.R. (1986). *The Economics of Public Utility Regulation*. Macmillan Press.

Cubbin, J. and Tzanidakis, G. (1998). Regression versus data envelopment analysis for efficiency measurement: an application to the England and Wales regulated water industry. ***Utilities Policy 7***, 75-85.

Debreu, G. (1951). The Coefficient of Resource Utilisation. ***Econometrica 19***, 225-234.

Decree on Price Determination of Communal Services (2005) [Uredba o oblikovanju cen komunalnih storitev]. Official Gazette of the Republic of Slovenia 45/2005, Ljubljana.

Denny, M. and Fuss, M. (1977). The Use of Approximation Analysis to Test for Separability and the Existence of Consistent Aggregates. ***American Economic Review 67(3)***, 404-418.

Deprins, D. and Simar, L. (1989a). Estimation de Frontières Déterministes avec Facteurs Exogènes d'Inefficacité. ***Annales d'Economie et de Statistique 14***, 117-150.

Deprins, D. and Simar, L. (1989b). Estimating Technical Inefficiencies with Corrections for Environmental Conditions with an Application to Railway Companies. ***Annals of Public and Cooperative Economics 60(1)***, 81-102.

Diewert, W.E. (1971). An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function. ***Journal of Political Economy 79(3)***, 481-507.

Diewert, W.E. and Wales, T.J. (1987). Flexible Functional Forms and Global Curvature Conditions. ***Econometrica 55***, 43-68.

Directive 2000/60/EC of the European Parliament and the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. OJ L 327.

Energy Agency of the Republic of Slovenia (AERS) (2004). Report on the Energy Sector 2003. Maribor.

Environment Protection Act (1993) [Zakon o varstvu okolja – ZVO]. Official Gazette of the Republic of Slovenia 32/1993, Ljubljana, valid until 2004.

Environment Protection Act (2004) [Zakon o varstvu okolja – ZVO-1]. Official Gazette of the Republic of Slovenia 41/2004, Ljubljana.

Estache, A. and Rossi, M.A. (2000). Comparing the Performance of Public and Private Water Companies in Asia and Pacific Region: What a Stochastic Costs Frontier Shows. Working Paper, Economic Development Institute, World Bank, Washington, D.C.

Estache, A., Rossi, M.A., Ruzzier, C.A. (2004). The Case of International Coordination of Electricity Regulation: Evidence from Measurement of Efficiency in South America. ***Journal of Regulatory Economics 25(3)***, 271-295.

Evans, D.S. and Heckman, J.J. (1984). A Test for Subadditivity of the Cost Function with an Application to the Bell System. ***American Economic Review 74***, 615-623.

Evans, R.G. (1971). "Behavioural Cost Functions for Hospitals. ***The Canadian Journal of Economics 4(2)***, 198-215.

Fabbri, P. and Fraquelli, G (2000). Costs and Structure of Technology in the Italian Water Industry. ***Empirica 27***, 65-82.

Färe, R., Grosskopf, S., and Lovell, C.A.K. (1985). *The Measurement of the Efficiency of Production.* Boston (MA), Kluwer Academic Publishers.

Färe, R., Grosskopf, S., and Lovell, C.A.K. (1988). Scale Elasticity and Scale Efficiency. ***Journal of Institutional and Theoretical Economics 144(4)***, 721-729.

Farrell, M.J. (1957). The Measurement of Productive Efficiency. ***Journal of the Royal Statistical Society, Series A, 120***, 253-281.

Farsi, M. and Filippini, M. (2004). Regulation and measuring cost efficiency with panel data models: application to electricity distribution utilities. ***The Review of Industrial Organization 25(1)***, 1-19.

Farsi, M., Filippini, M. and Greene, W. (2005). Efficiency Measurement in Network Industries: Application to the Swiss Railway Companies. ***Journal of Regulatory Economics 28(1)***, 69-90.

Farsi, M., Fillipini, M. and Kuenzle, M. (2005). Unobserved heterogeneity in stochastic cost frontier models: an application to Swiss nursing homes. ***Applied Economics 37(18)***, 2127-2141.

Farsi, M., Fillipini, M. and Kuenzle, M. (2006). Cost Efficiency in Regional Bus Companies: An Application of Alternative Stochastic Frontier Models. ***Journal of Transport Economics and Policy 40(1)***, 95-118.

Feigenbaum, S. and Teeples, R. (1983). Public Versus Private Water Delivery: A Hedonic Cost Approach. ***The Review of Economics and Statistics 65(4)***, 672-678.

Filippini, M. (1996). Economies of scale and utilization in the Swiss electric power distribution industry. ***Applied Economics 28***, 543– 550.

Filippini, M., Farsi, M., Fetz, A. (2005). Benchmarking Analysis in Electricity Distribution. CEPE Report Nr. 4.

Fuss, M., McFadden, D., Mundlak, Y. (1978). *A Survey of Functional Forms in the Economic Analysis of Production*. In Fuss, M. and McFadden, D. (eds.). *Production Economics: A Dual Approach to Theory and Applications*. New York, North-Holland, 219-268.

Gallant, A.R. (1981). On the Bias in Flexible Functional Forms and Essentially Unbiased Form. ***Journal of Econometrics 15***, 211-245.

Garcia, S. and Thomas, A. (2001). The Structure of Municipal Water Supply Costs: Application to a Panel of French Local Communities. ***Journal of Productivity Analysis 16***, 5-29.

Guyomard, H., Vermersch, D. (1989). Derivation of long-run factor demands from short-run responses. ***Agricultural Economics 3***, 213–230.

Greene, W.H. (1980a). Maximum Likelihood Estimation of Econometric Frontier Functions. ***Journal of Econometrics 13(1)***, 27-56.

Greene, W.H. (1980b). On the Estimation of a Flexible Frontier Production Model. ***Journal of Econometrics 13(1)***, 101-115.

Greene, W.H. (1990). A Gamma Distributed Stochastic Frontier Model. ***Journal of Econometrics 46(1/2)***, 141-164.

Greene, W.H. (1997). *Frontier Production Functions*, In Pesaran, M.H. and Schmidt, P. (eds.). *Handbook of Applied Econometrics, Volume II: Microeconomics.* Blackwell Publishers Ltd, 81-166.

Greene, W.H. (2000). *Econometric Analysis*, 4th ed. New Jersey, Prentice Hall International, Inc.

Greene, W.H. (2001). Estimating Econometric Models with Fixed Effects. Working Paper, Department of Economics, Stern School of Business, New York University.

Greene, W.H. (2002a). Alternative Panel Data Estimators for Stochastic Frontier Models. Working Paper, Department of Economics, Stern School of Business, New York University.

Greene, W.H. (2002b). Fixed and Random Effects in Stochastic Frontier Models. Working Paper, Department of Economics, Stern School of Business, New York University.

Greene, W.H. (2003). Distinguishing Between Heterogeneity and Inefficiency: Stochastic Frontier Analysis of the World Health Organization's Panel Data on National Health Care Systems. Working Paper, Department of Economics, Stern School of Business, New York University.

Greene, W.H. (2005). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. ***Journal of Econometrics 126***, 269-303.

Griffin, R.C., Montgomery, J.M., Rister, M.E. (1987). Selecting Functional Form in Production Function Analysis. ***Western Journal of Agricultural Economics 12(2)***, 216-227.

Hall, G.R. (2000). Analysis of Alternative Ratemaking Methodologies. Asian Development Bank, The Energy Regulatory Board, and Department for Energy. Workshop Paper No. 2.

Hall, R.E. (1973). The Specification of Technology with Several Kinds of Output. ***Journal of Political Economy 81(4)***, 878-892.

Hill, L.J. (1995). A Primer on Incentive Regulation for Electric Utilities. Oak Ridge National Laboratory, ORNL/CON-422, Tennessee.

Hrovatin, N. and Bailey, S.J. (2001). Implementing the European Commission's water pricing communication: cross-country perspectives. ***Utilities Policy 10***, 13-24.

Hrovatin, N. (ed.) (2002). *Strategy of the Local Public Utilities' Development in Slovenia* [*Strategija razvoja lokalnih gospodarskih javnih služb v Sloveniji*]. Svetovalni center, Ljubljana.

Hsiao, C. (2003). *Analysis of Panel Data*, 2nd Edition. Cambridge University Press.

Huang, C.J. and Liu, J.-T. (1994). Estimation of a Non-Neutral Stochastic Frontier Production Function. ***Journal of Productivity Analysis 5(2)***, 171-180.

Jamasb, T. and Pollitt, M. (2001). Benchmarking and regulation: international electricity experience. ***Utilities Policy 9***, 107-130.

Jehle, G.A. and Reny, P.J. (1998). *Advanced Microeconomic Theory*. Addison-Wesley.

Jondrow, J., Lovell, C.A.K., Materov, I.S., Schmidt, P. (1982). On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model. ***Journal of Econometrics 19(2/3)***, 233-238.

Jorgenson, D.W. and Griliches, Z. (1967). The Explanation of Productivity Change. ***Review of Economic Studies 34(3)***, 249-283.

Joskow, P.J. and Schmalensee, R. (1986). Incentive regulation for electric utilities. ***Yale Journal on Regulation 4***, 1-49.

Kavčič, S: (2000). Effect of Depreciation on the Performance of Communal Utilities – Research Results [Vpliv amortizacije na poslovanje komunalnih podjetij – Predstavitev raziskovalne naloge]. Svetovalni center, Ljubljana.

Kennedy, D. (1997). Competition in the Water Industry. Discussion Paper 16, Chartered Institute of Public Finance and Accountancy, Centre for the Study of Regulated Industries, London.

Kim, H.Y. and Clark, R.M. (1988). Economies of Scale and Scope in Water Supply. ***Regional Science and Urban Economics 18***, 479-502.

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 13. New York, Wiley.

Kumbhakar, S. C. (1990). Production Frontiers, Panel Data and Time-Varying Technical Inefficiency. ***Journal of Econometrics 46(1/2)***, 201-212.

Kumbhakar, S.C., Ghost, S. and McGuckin, J.T. (1991). A Generalized Production Frontier Approach for Estimating Determinants of Inefficiency in US Dairy Farms. ***Journal of Business and Economic Statistics 9(3)***, 279-286.

Kumbhakar, S. C. and Lovell, C.A.K. (2000). *Stochastic Frontier Analysis*. Cambridge University Press.

Laffont, J.J. and Tirole, J. (1993). *A Theory of Incentives in Procurement and Regulation*. Cambridge, MIT Press.

Lancaster, K.J. (1966). New Approach to Consumer Theory. ***Journal of Political Economy 74***, 132-157.

Lau, L.J. (1974). *Comments*. In Intriligator, M.D. and Kendrick, D.A. (eds.). *Frontiers of Quantitative Economics, Volume II*. Amsterdam, North-Holland Publishing Company, 176-199.

Lau, L.J. (1986). *Functional Forms in Econometric Model Building*. In Griliches, Z. and Intriligator, M.D. (eds.). *Handbook of Econometrics, Volume III*. Amsterdam, North-Holland Publishing Company, 1515-1565.

Law on Financing Municipalities (1994) [Zakon o financiranju občin – ZFO]. Official Gazette of the Republic of Slovenia 80/1994, Ljubljana.

Law on Local Self-Government (1993) [Zakon o lokalni samoupravi – ZLS]. Official Gazette of the Republic of Slovenia 72/1993, Ljubljana.

Law on Prices (1991) [Zakon o cenah – ZCen]. Official Gazette of the Republic of Slovenia/I, 1/1991, Ljubljana, valid until 1999.

Law on Price Control (1999) [Zakon o kontroli cen – ZKC]. Official Gazette of the Republic of Slovenia, 63/1999, Ljubljana.

Lee, Y.H. and Schmidt, P. (1993). *A Production Frontier Model with Flexible Temporal Variation in Technical Inefficiency*. In Fried, H.O., Lovell, C.A.K. and Schmidt, S.S. (eds.). *The Measurement of Productive Efficiency: Techniques and Applications*. New York, Oxford University Press, 237-255.

Lewis, T.R. and Sappington, D.E.M. (1989). Regulatory Options and Price-Cap Regulation. ***Rand Journal of Economics 20(3)***, 405-416.

Littlechild, S. (1983). Regulation of British Telecommunications' Profitability. HMSO, London.

Maddala, G.S. (2001). *Introduction to Econometrics*, 3rd ed. John Wiley & Sons, Ltd.

Mann, P.C. (1993). Water-Utility Regulation: Rates and Cost Recovery. Policy Study Nr.155, Reason Public Policy Institute, US.

Mas-Colell, A., Whinston, M.D., Green, J.R. (1995). *Microeconomic Theory*. 2nd ed. Oxford University Press.

Massarutto, A. (1999). Water management and water prices in Italy. Paper presented at the conference Pricing Water: Economics, Environment & Society, Sintra, Portugal, 6–7 September 1999. European Commission DG XI, Brussels and Instituto da Agua, Portugal.

McFadden, D. (1978). *Cost, Revenue and Profit Functions*. In Fuss, M. and McFadden, D. (eds.). *Production Economics: A Dual Approach to Theory and Applications*. New York, North-Holland, 3-109.

Meeusen, W. and van den Broeck, J. (1977). Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. ***International Economic Review 18(2)***, 435-444.

Mizutani, F. and Urakami, T. (2001). Identifying network density and scale economies for Japanese water supply organizations. ***Papers in Regional Science 80***, 211-230.

Morishima, N. (1967). A Few Suggestions on the Theory of Elasticity. ***Economic Review 16*** (in Japanese), 73-83.

Mrak, M. (1997). Infrastructure Investment Needs in Slovenia. ***IB Revija 31(12)***, 9-23.

Mrak, M. (2000). Communal Infrastructure in Slovenia: Survey on Investment Needs and Policies Aimed at Encouraging Private Sector Particiation. World Bank Technical Paper No. 483, Europe and Central Asia Poverty Reduction and Economic Management Series, IBRD, World Bank, Washington, D.C.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross-Section Data. ***Econometrica 46(1)***, 69-85.

Netz, J.S. (1999). Price Regulation: A (Non-Technical) Overview. Department of Economics, Purdue University.

OFWAT (1993). Setting Price Limits for Water and Sewerage Services. The framework and approach to the 1994 Periodic Review. Birmingham, UK.

OFWAT (1998). Assessing the scope for future improvements in water company efficiency: A technical paper. Birmingham, UK.

OFWAT (1999). Final Determinations: Future water and sewerage charges 2000-05. 1998 Periodic Review. Birmingham, UK.

OFWAT (2004). Future water and sewerage charges 2005-10: Final determinations. 2004 Periodic Review. Birmingham, UK.

Orea, L. (2002). A Parametric Decomposition of a Generalized Malmquist Productivity Index. ***Journal of Productivity Analysis 18(1)***, 5-22.

Oum, T.H. and Tretheway, M.W. (1989). Hedonic vs. General Specifications of the Translog Cost Function. ***The Logistics and Transportation Review 25(1)***, 3-21.

Panzar, J.C. and Willig, R.D. (1977). Economies of Scale in Multi-Output Production. ***Quarterly Journal of Economics 91(3)***, 481-493.

Panzar, J.C. and Willig, R.D. (1981). Economies of Scope. ***American Economic Review 71***, 268-272.

PASEF (2004). Data base of final accounts of Slovenian companies (1994-2004). Faculty of Economics, Ljubljana.

Pitt, M. and Lee, L.F. (1981). The Measurement and Sources of Technical Inefficiency in the Indonesian Weaving Industry. ***Journal of Development Economics 9***, 43-64.

Polachek, S. and Yoon, B. (1994). Estimating a Two-Tiered Earnings Function. Working Paper, Department of Economics, State University of New York, Binghamton.

Polachek, S. and Yoon, B. (1996). Panel Estimates of a Two-Tiered Earnings Frontier. ***Journal of Applied Econometrics 11***, 169-178.

Public Utilities Act (1993) [Zakon o gospodrskih javnih službah – ZGJS]. Official Gazette of the Republic of Slovenia 32/1993, Ljubljana.

Pulley, L.B. and Braunstein, Y.M. (1992). A composite cost function for multiproduct firms with an application to economies of scope in banking. ***Review of Economics and Statistics 74(2)***, 221-230.

Reifschneider, D. and Stevenson, R. (1991). Systematic Departures from the Frontier: A Framework for the Analysis of Firm Inefficiency. ***International Economic Review 32(3)***, 715-723.

Richmond, J. (1974). Estimating the Efficiency of Production. ***International Economic Review 15(2)***, 515-521.

Roberts, M.J. (1986). Economies of Density and Size in the Production and Delivery of Electric Power. ***Land Economics 62(4)***, 378-387.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. ***Journal of Political Economy 82(1)***, 34-55.

Rules on Price Determination of Obligatory Local Public Utilities for Environment Protection (2004) [Pravilnik o oblikovanju cen storitev obveznih občinskih gospodarskih javnih služb varstva okolja]. Official Gazette of the Republic of Slovenia 128/2004, Ljubljana.

Ryan, D.L. and Wales, T.J. (2000). Imposing local concavity in the translog and generalized Leontief cost functions. ***Economic Letters 67***, 253-260.

Schmidt, P. and Sickles, R.C. (1984). Production Frontiers and Panel Data. ***Journal of Business and Economic Statistics 2(4)***, 367-374.

Salvanes, K.G. and Tjøtta, S. (1998). A Note on the Importance of Testing for Regularities for Estimated Flexible Functional Forms. ***Journal of Productivity Analysis 9***, 133-143.

Shephard, R.W. (1953). *Cost and Production Functions*. Princeton, Princeton University Press.

Shephard, R.W. (1970). *The Theory of Cost and Production Functions*. Princeton, Princeton University Press.

Shleifer, A. (1985). A Theory of Yardstick Competition. ***Rand Journal of Economics 16***, 319-327.

Sibley, D. (1989). Asymmetric Information, Incentives and Price-Cap Regulation. ***Rand Journal of Economics 20(3)***, 392-404.

Simar, L., Lovell, C.A.K. and Eeckaut, P. Vanden (1994). Stochastic Frontiers Incorporating Exogenous Influences on Efficiency. Discussion Paper No.9403, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Slovenian Accounting Standards (2004) [Slovenski računovodski standardi – SRS]. The Slovenian Institute of Auditors, Ljubljana.

Spady, R.H. and Friedlaender, A.F. (1978). Hedonic Cost Functions for the Regulated Trucking Industry. ***The Bell Journal of Economics 9(1)***, 159-179.

Statistical Yearbook of the Republic of Slovenia 2001 (2001). Statistical Office of the Republic of Slovenia – SURS, Ljubljana.

Statistical Yearbook of the Republic of Slovenia 2003 (2003). Statistical Office of the Republic of Slovenia – SURS, Ljubljana.

Stevenson, R.E. (1980). Likelihood Functions for Generalized Stochastic Frontier Estimation. ***Journal of Econometrics 13(1)***, 57-66.

Stewart, M. (1993). Modelling Water Costs 1992-93: Further research into the impact of operating conditions on company costs. Ofwat Research Paper Number 2. OFWAT, Birmingham, UK.

Svetovalni center (1997). Analysis of Communal Sector Prices in the period 1991-1997. Ljubljana.

Svetovalni center (2001). Analysis of Communal Sector Prices in 1999 and 2000. Ljubljana.

Štruc, M. (ed.) (1997). *Strategy of the Communal Sector Development in the Republic of Slovenia* [*Strategija razvoja komunalnega gospodarstva v Republiki Sloveniji*]. Svetovalni center, Ljubljana.

Teeples, R. and Glyer, D. (1987). Cost of Water Delivery Systems: Specification and Ownership Effects. ***Review of Economics and Statistics 69***, 399-407.

Thanassoulis, E. (2000). The use of data envelopment analysis in the regulation of UK water utilities: Water distribution. ***European Journal of Operational Research 126***, 436-453.

Uzawa, H. (1982). Production Functions with Constant Elasticities of Substitution. ***Review of Economic Studies 29***, 291-299.

Varian, H. (1984). *Microeconomic Analysis*. 2nd Edition. New York, W.W. Norton.

Vass, P. (2000): Global issues, Utility regulation in Central and Eastern Europe: economic regulation of privatised utilities and network industries, ***Utility regulation 2000 3***, 5-8.

Vickers, J. and Yarrow, G. (1988). *Privatisation: An Economic Analysis*. MIT Press, Cambridge, Mass.

Wang, H.-J. (2002). Heteroscedasticity and Non-Monotonic Efficiency Effects of a Stochastic Frontier Model. ***Journal of Productivity Analysis 18***, 241-253.

Wang, H. and Schmidt, P. (2002). One-Step and Two-Step Estimation of the Effects of Exogenous Variables on Technical Efficiency Levels. ***Journal of Productivity Analysis 18***, 129-144.

Water Act (1981) [Zakon o vodah – ZV]. Official Gazette of the Socialistic Republic of Slovenia 38/1981, Ljubljana, valid until 2002.

Water Act (2002) [Zakon o vodah – ZV-1]. Official Gazette of the Republic of Slovenia 67/2002, as amended by 41/2004, Ljubljana.

White, H. (1980). Using Least Squares to Approximate Unknown Regression Functions. ***International Economic Review 21(1)***, 149-170.

Williamson, B. and Toft, S. (2001). The Appropriate Role of Yardstick Methods in Regulation. National Economic Research Associates (NERA), London, UK.

Winsten, C.B. (1957). Discussion on Mr. Farrell's Paper. ***Journal of Royal Statistical Society Series, Series A (General) 120(3)***, 282-284.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. ***Journal of the American Statistical Association 58***, 977-992.

Zhu, J. (2003). *Quantitative Models for Performance evaluation and Benchmarking, Data Envelopment Analysis with Spreadsheets and DEA Excel Solver*. Kluwer Academic Publishers.