PRINCIPLES OF PRODUCTIVITY MEASUREMENT

AN ELEMENTARY INTRODUCTION TO QUANTITATIVE RESEARCH ON THE PRODUCTIVITY, EFFICIENCY, EFFECTIVENESS AND QUALITY OF THE PUBLIC SECTOR (3rd EDITION)

JOS L.T. BLANK
VIVIAN G. VALDMANIS

ISBN 978-90-827258-0-3

Stichting IPSE Studies
Delft, Netherlands
info@ipsestudies.nl
j.blank@ipsestudies.nl
http:// www.ipsestudies.nl

Contents

Preface

The public sector contributes considerably to social welfare. Education, law enforcement and healthcare are important sectors for a smoothly functioning economy contributing to a socially just society. Because these sectors are often financed by taxes and premiums and because the market's regulation may be lacking, insight into the performances of these facilities is of utmost importance. Analyzing the productivity, effectiveness and efficiency of public services has therefore become imperative. Over the last 25 years the academic developments in this field have been most impressive. For policy makers and managers of public firms the advanced methods are, however, not quite as accessible as possible. The acceptance of research is therefore often negligible. This book attempts to bridge the gap between science and practice. Theoretical concepts will be explained in a comprehensible manner. With the help of practical examples we will explain why academic methods should be preferred to numerous heuristic methods. Results from research for policy makers will only be used if the work is understandable and appropriate for the policy question being addressed and can result in actionable decisions.

For the realization of this book we have combined more than sixty years of research experience in this field. Insights have been realized by the many interesting and profound discussions with colleagues and fellow academics. We are most appreciative for that. Naturally we are responsible for the content of this book. Discussions will always take place, both policy-wise and academic. This of course means that this book requires a regular update. Therefore we made a thorough revision of our book, published in 2013. However the research proceeds and therefore we invite the reader to comment, enter into discussion and make suggestions for identifying changes in the public sector and new calls for quantitative analysis. We also include current websites and video websites that discuss statistical approaches. So we published a complete revised book in 2017. However, the 2017 issue contained some disturbing errors. We corrected these errors and now we issued a third version in 2019.

One final introductory note. In order to preserve consistency we use the metric system to denote spatial size (hectares and meters); and we use the euro as the unit of currency.

Lastly we would like to say thank you to Thomas Niaounakis for all his help with the editing of the book.

Jos Blank,
Innovation and Public Sector Efficiency Studies, Delft, The Netherlands

Vivian Valdmanis,
Western Michigan University, United States of America

September 2019

1 Performances of the public sector

1.1 Toward a productive public sector

During the last two decades there has been a growing interest in the functioning of the public sector. One of the reasons is the profound influence the public sector has on the economy as a whole. Decent law enforcement, a well-trained and healthy labor force and adequate infrastructure are important assets for a smoothly functioning economy. Another reason for the increasing need for public services is when the private markets do not adequately provide enough services to supply social needs. Even if motivated by social justice, there are not the economic incentives for the private for-profit sector to engage in these markets. Hence, the public sector needs to have the larger role in the production of social/public goods.

The drawback of a sizeable public sector is that these services are largely financed by collective resources. It is therefore important to set up the public sector as productive as possible in order to achieve efficient and innovative behavior.

Apart from an efficient execution of public services, it is also important to assess whether the public sector, in fact, provides the services that contribute to achieving substantial social goals. For example, it is questionable whether issuing large amounts of fines by the police, however efficiently executed, still contributes to social safety. The content of education in relation to the opportunities in the labor market and social participation is another example. Social welfare, such as the right to decent healthcare for each citizen plays an important part which given the aging population will lead to an ever expanding healthcare system. The accessibility for the group of elderly people or poorer people is often motivated by social justice reasons, especially if services for these individuals are not profitable, i.e., higher cost patients or clients to treat. If social insurance, health insurance, long term care to be provided by the private sector prices may become prohibitive thereby limiting needed access in these markets. Measuring the efficiency of public services is central in posing these types of questions. However, responsible spending by the government requires posing the questions of efficiency and productivity.

Effectiveness and efficiency of public services is concentrated in politics and policy. Many policy questions relate to the effects of how things are managed and financed, how capacity is planned, and how the effectiveness and efficiency of public services are monitored. Because of decentralization of responsibilities by privatizations, liberalization of markets and deregulation, the interest also comes from presidents/leaders of public firms. Instead of being the executors of legislation they have become more like entrepreneurs, who set up their own management and must make their own strategic choices. Given current austerity measures, the strains on all levels of government budgets, and the increasing importance of minimizing cost increases the demand for more research and appropriate application of quantitative methods.

Even though decision makers and analysts without formal training in traditional economics who are pursuing careers or advancement in public policy, they can profit from understanding the methods and implications well enough to use the findings in discerning appropriate, effective courses of action to meet specified objective.

Therefore, some public sector researchers have focused on the functioning of the public sector, and based their assessments on effectiveness and efficiency. There is a literature in public administration available on new public management (see e.g. McLaughlin et al., 2002).

We can say a great deal about performance measurement in the public sector. Advocates of rigorous policy study indicate that performance measurement contributes to more rational contemplations in the management of public services. Critics, however, stress that it might be impossible to accurately measure various public values, and also point at the negative effects of performance measurement due to strategic behavior. In a standard book of performance measurement in the public sector De Bruijn (2007) elaborately discusses all these aspects. He depicts a balanced image of all advantages and disadvantages of performance measurement. He especially focuses on the dynamic aspects of performance measurement behavior, the tendency toward collecting data, and the collective fixation on figures. One way to reconcile the advocates and the critics is to design research that is based on best practice. This approach demonstrates relative efficiency so that decision-makers can ascertain what improvements in efficiency or productivity are possible in their organization given other organizations in similar operating environments and not on a perceived unattainable outcome.

The increased attention for the subject has been translated into empirical research of the efficiency of public services. Researchers have developed and presented policy based results on these so-called benchmarks. Comparing various aspects of firms, has become generally accepted in many sectors, as for example, in the approach by Kaplan & Norton (1996) who developed the Balanced Scorecard (BSC). This is a map with which an organization is illustrated from four different perspectives: finances, customer satisfaction, internal processes and the ability to learn.

More useful and advanced alternatives, that in concept are more effective, have been developed. Important techniques, such as the stochastic frontier analysis (SFA) and the data envelopment analysis (DEA) have been developed to measure productivity in an integral and coherent way. It is striking that these techniques are not only applied to public sectors such as healthcare and education but also to private sectors such as banking and agriculture.

1.2 Government policy

In the introduction, we have already identified the shift of responsibilities from being government bureaucrats to becoming executors. Examples of government services that have been converted into independent managing bodies include financing that have been adjusted, markets that have been liberalized and several other forms of deregulation. Most changes were especially motivated by supposed considerations of efficiency. These independent bodies, where the most knowledge about local circumstances is accumulated, can deal more efficiently with their responsibilities, which lead to a better allocation of resources. Shifting the responsibilities alone could also be a vital stimulus to work as efficiently as possible.

It must be reiterated, that the market chooses areas wherein profits can be garnered. Simply posing markets in social areas have not been made based on strong academic substantiation, but rather on ideological considerations. Self-interest, political expediency, and preferring the status quo to change are often seen as significant contributions as to why policy makers opt out of accepting research that would substantiate ways to increase social welfare rather than arguing for markets without substantial evidence.

An effective and efficient execution of public services guaranteeing a number of public values, based on politically formulated goals as the starting point, benefits from a more rational approach. By using theories from economics, market organization, and regulation measuring effectiveness and reporting on effectiveness can be achieved.

1.3 Management policy

In regulated public services, directed by the government, the management of a public firm is trying to correctly enforce the rules. Most notably the natural monopoly argument within economics requires only one firm provide a service, but with regulated fees and restrictions that would in fact “mimic” what prices would be in a competitive market. Within decentralized sectors, the management must focus on making decisions about business economics and strategy demanding other skills of managers. Solid decision-making by management must have access to relevant and transparent information. This requires a different organization and structure of the internal information service. Therefore, there was such a call for increased regulation as a result of the meltdown on Wall Street in 2007-2008, where such a lack of transparency and accurate information on exotic trades and sub-prime housing mortgages went either unnoticed or was met with complacency.

If there is no accurate and adequate information available regulations cannot be implemented to induce private markets to operate efficiently. It is up to the public sector to implement regulations again to correct the market failures such as asymmetric information, inadequate pricing, and lack of competition – all that could doom an ideal market from operating.

1.3.1 The chaff and the wheat

With the increasing strategic interest for more effective public policies and administration, the need for more research has also grown. Starting in the late 1990’s, benchmarking or supplying market based information has become very popular. Based on numerous index numbers firms are compared to one another. Conclusions from this research are often controversial, because the coherency between the different aspects is frequently disregarded. More often than not, strongly normative assumptions are made about what is a best practice. In other cases the quality of data has not improved, but rather deteriorated thereby reducing the ability to provide accurate insights for management.

1.4 This Book

1.4.1 Purpose

In this chapter's introduction the considerable importance of effective and efficient public services (with some examples) was identified. In order to realize this insight into the functioning of the system and institutions is of fundamental importance. This takes adequate quantitative research that meets today's academic standards. This is the major motivation in writing this book: making academic research methods of the public sector's functioning accessible to the general public.

The purpose of this book is to inform the reader of the backgrounds of measuring important notions from public policy analysis, such as productivity, effectiveness and efficiency. Important questions include: what do these concepts mean, how do you measure them, and how do you make these concepts ready for use in concrete public services? In fact, the book is an elementary introduction into the specialty of integral measurement of productivity, effectiveness and efficiency.

Many policy reports contain numerous documents and empirical research results using these kinds of indicators. Whether or not these results are all reliable, how they should be interpreted, and whether the employed academic method is justified are legitimate questions. The content of this book assists the reader in understanding and critically assessing these kinds of analyses.

Furthermore, this book also provides examples for research in various fields. It contains references of applied research in several public sectors, such as healthcare, education, and police. In some cases we give details about a relevant application (presented in tables) and in others we only summarize the findings and results.

By making research results more accessible for a general public the book is also designed to encourage more and better designed research. This will contribute to a better organized public sector by means of a more substantial economic rationalization of government policy (finance, regulation), and by improving corporate strategies, corporate models and management.

1.4.2 Who makes up the public sector and how do they contribute to social welfare?

In this book, we focus on problems in the public sector, such as hospitals, public health, public transport and schools. These are sectors that do not have fully transparent markets, where there is profound regulation because of safeguarding public values, and where a diffuse product is supplied. This requires special research techniques. These techniques, however, are not exclusively for analyzing public sectors. This type of research is also popular in firms with many independently functioning units, such as banks. It is also realistic that the measurement techniques that we will discuss are applied in a strictly technical context. An example is the assessment of technical specifications in order to find out which type of engine performs best, i.e., the evaluation is done in the engineering sense.

This book will not train the reader to become a full-fledged researcher in this field. For that other excellent books and/or analytical surveys are available (see e.g. Balk, 2003; Blank & Valdmanis, 2008; Blank & Lovell, 2000; Coelli et al., 2005; Cooper et al., 2007; Färe & Primont, 1995; Hulten et al., 2001; Kumbhakar & Lovell, 2000).

1.4.3 Who should be interested in this book?

This book is meant for everyone who wishes the public sector well and who are interested in how the public sector in general and public firms in particular perform. Policy makers at departments and municipalities, politicians, representatives of umbrella organizations, unions, consumer organizations, members of boards, supervisors, managers of firms and researchers and students in the field of public administration, public management, public policy analysis, and public finances can all glean something from this book. It is also meant to spark interest for future researchers who wish to conduct the types of analyses presented here especially as dynamic changes may be in the offing requiring solid research based on theoretical grounds.

As much as possible the complex statistics and mathematics that are common in productivity measurement will be avoided. The theory is explained based on simple style examples and comprehendible figures. Some insight into basic mathematical equations or being able to grasp graphics is essential though.

1.6 Bookmark anchor

Chapter 2 contains an overview and explanation of the major concepts and definitions, such as productivity, effectiveness, and efficiency. This chapter also comprises an explanation of these concepts, such as resources, intermediary products, final products, quality, and effects.

In Chapter 3 the microeconomic theory is described. Here, we discuss the various economic models, of which the formulation is determined by assumptions about the economic behavior of firms and the constraints. These economic models are especially important because they offer a fixed point to evaluate firms. It is clear that the trading results of a non-profit firm are different from those of a firm aiming to maximize its profits.

Data collection is the focus of Chapter 4 including the way data should be gathered, checking the plausibility of the data, and missing data imputation techniques.

We explain how measuring productivity, effectiveness, and efficiency with the available data is accomplished in Chapter 5. The chapter will delve into the difference between partial and integral methods and will show that the latter is far better. As part of this approach, we explain the measurement of SFA, DEA and semi-parametric methods.

The development of research of productivity, effectiveness and efficiency is described in Chapter 6. It explains about setting up an economic (reference) model and the definition of the various variables.

In Chapter 7 we focus on policy evaluations where the outcome of the productivity analysis must be perceived in consistency with the policy instruments. Different variants, such as the effect of market forces on the productivity of healthcare will be discussed.

Analyses of management and productivity are covered in Chapter 8. Central to this them include the influence of management, corporate models and corporate strategy on a firm's productivity. We especially focus on the effects of innovation.

In Chapter 9 we conclude with commentary on the importance of quantitative research wherein truth can be spoken to power.

Throughout the book, specific examples from the literature will be given with the concepts and approaches illustrated in each chapter.

2 Relevant definitions

2.1 Introduction

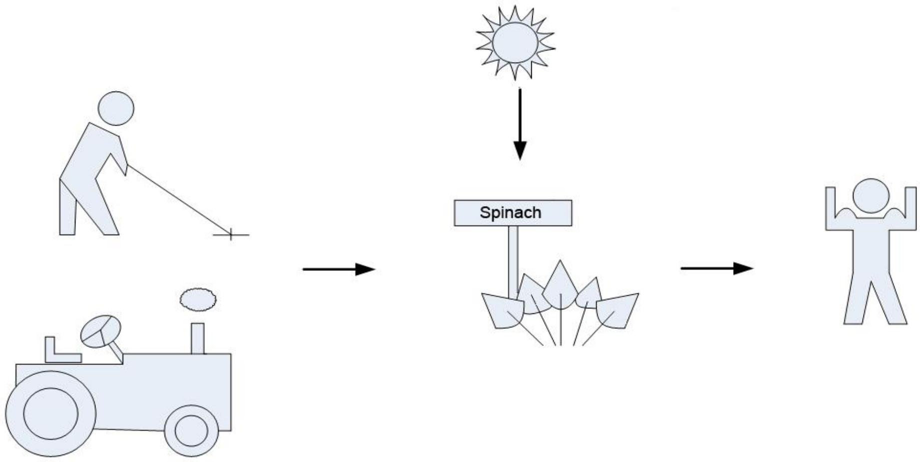
Confusion can arise when measuring performance of public services. Some decision-makers/researchers define performance as a measurement of services provided, such as the number of patients treated in a hospital or the number of a school's students who pass their exams. Others define performance as a measurement of the relation between the provided services and the efforts required to provide those services, such as the number of personnel.

In order to avoid confusion, this chapter deals with the major relevant definitions and sheds light on the relations among them. The concept of performance is avoided as much as possible in the book. Instead, the focus is on more specific definitions, namely:

- Process;
- Productivity;
- Efficiency;
- Effectiveness
- Environmental influences;
- Levels of research.

First, we discuss process and productivity. Figure 2.1 illustrates the evaluation of the functioning of public services. The process starts with the employment of resources, such as personnel (the farmer) and capital (the rake and the tractor). These will be converted into products and/or services, in this case spinach. This conversion is influenced by environmental factors, such as the number of hours of sunshine. The products and services are consumed and contribute to a social effect, namely, an outcome. The relation between the employment of resources and the output determines the productivity. The relation between the employment of resources and the outcome determines the effectiveness. In some cases when outcomes cannot be or are not observed, outputs may be used if they can be theoretically linked to outcomes. In other words, productivity relates inputs to the production of spinach; efficiency demonstrates how using spinach as an input results in an outcome defined as healthier diets for those who eat spinach.

Figure 2-1 Relation between resources, environment, output and outcome



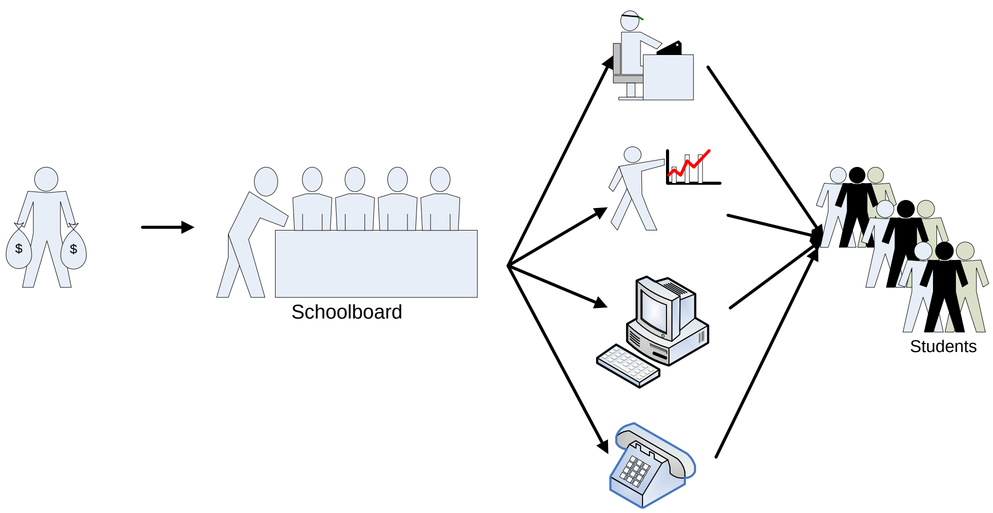
Even though this appears to be a simple example, Figure 2-1 is in fact too extensive. Everything in the productivity process can be conceptually reduced to the definitions of output and resources. The environment can be regarded as a resource over which the actors have no control. The number of hours of sunshine -at a given location- cannot be influenced, but it is one of the most vital ingredients that make spinach grow. An effect can also be interpreted as a form of output, but it is measured in other units, such as health, social participation, and self-help.

Another commonly used concept is the output's quality. In Figure 2-1 quality can be integrated by differentiating between fresh and shriveled spinach. We now have two products. If the market is operating efficiently, farmers who produce shriveled spinach will not be in business for long, if there is only demand for fresh spinach. But in the case of the production of social goods, the concept of quality is far more subtle. Hence, the need exists for government regulation to ensure minimum as well as superior quality, if quality or appropriateness cannot be adequately measured by consumers in the market. Otherwise, there will be a significant welfare loss, such as paying too much for minimum quality or not paying enough for superior quality. This is dealt with further in later chapters.

Because the discussed subjects mostly take place in an economic context, price (or relative appreciation), economic goal and economic precondition are three key phrases. The entire concept can be reduced to these elementary parts.

Figure 2-1 is a purely technical relation, with which resources such as personnel, capital, and material are converted into products (or outcomes). When determining the relation between the resources and that between the different products, both economic motives and political considerations play a role. Turning to the example of schools, this more complicated structure is illustrated in Figure 2-2.

Figure 2-2 Allocation of resources in a school at output maximization



The management has a budget and can utilize the money in the budget to hire teachers or support staff, as well to purchase computers, to hire a cleaning service, or to pay the electricity bill. The management must try to do this in such a way that as many students as possible leave school with a diploma. Again, a diploma is an output but the real value or outcome of a diploma is whether the students receiving it can either get admitted to a reputable university or obtain a job. Hence, some for-profit schools and universities apply market objectives but may or may not provide an adequate education (see the recent debate over for-profit educational institutions such as "Trump University") The for-profit higher education system has been a contentious issue in the US requiring governmental oversight. Similar schools may differ considerably in terms of results for a variety of reasons. Perhaps management does not quite know how to proportionally employ the resources, or perhaps it focuses on other goals that are not in accordance with the main goal of having students perform as well as possible. Without formal modelling, tying budgets together with output production without understanding the relation can lead to mistakes in decision making. Therefore, one must examine the formal underlying relationship converting inputs into outputs.

2.2 Productivity

Productivity is nothing more than the relation between supplied services and the resources necessary to produce them. When a firm supplies only one service (or output y) and the production requires only personnel (or input x), productivity equals:

$$\boldsymbol{Productivity} = \frac{y}{x} \tag{2-1}$$

The productivity of a park-keeping service, for example, equals the number of tilled square meters of public park per full-time employee of the park-keeping service. For this purpose, we shall disregard the resource 'hoe' as well as differences in the quality of work that may occur.

If we now compare the results of the various municipalities, we can determine how well one municipality performs vis-à-vis other municipalities, that is, the best practice approach described in Section 1.1. In this example, one must clarify that the concept of productivity is relative and only has meaning when the outcome of the above-mentioned equation can be compared with that of other firms.

If a firm or a governmental agency supplies more than one service and therefore it must also employ various resources rendering Equation (2-1) as no longer applicable. The productivity then equals the ratio between the volume of the output and the volume of the resources. A special case is the situation when prices for products and resources are available and when weights are utilized; then the productivity indicates the output per cost unit, which is also known as the profit ratio. The common equation for productivity is then:

$$Prod = \frac{p_1 y_1 + p_2 y_2 + \ldots + p_M y_M}{w_1 x_1 + v_2 x_2 + \ldots + w_N x_N} \tag{2-2}$$

$Prod$ = productivity;
p_m = weight of product m;
w_n = weight of resource n;
y_m = quantity of product m;
x_n = quantity of resource n.

In the example of the park-keeping service, this means that the workers not only till parks but also mow lawns. The supplied services now consist of the number of square meters tilled, and the number of square meters mowed. In the original example, the use of hoes was ignored for the sake of convenience. In the new example, the simplification does not apply, because the park-keeping service also employs motorized lawn mowers that, above all, need inputs such as petrol and maintenance. This adjusted example features multiple production and multiple resources (which are referred to as multiple-inputs, multiple-outputs) which is common in the social welfare context.

Box 2-1 Overview of important variables and symbols

For the sake of efficiency the rest of this book frequently features symbols. This box defines all the symbols and their meanings.

y_m = quantity of service m delivered (there are M services)
x_n = quantity of resource n used (there are N resources)
z_l = quantity of environmental aspect l (there are L environmental aspects)
p_m = price for service m to be paid by user (e.g. price per trip)
w_n = price for resource n to be paid (e.g. wages per hour)
R = total revenues of payment by users
C = total costs for all resources
Π = profit, being the difference between revenues and costs.

Please note that:
$R = p_1y_1 + .. + p_my_m$; is the sum of the revenue of all M services, when the quantity of each service is multiplied by its price;
$C = w_1x_1 + \cdots + w_nx_n$; is the sum of the costs of all resources, when the quantity of each resource is multiplied by its price;
$\Pi = R - C$; is profit, namely the difference between revenue and costs.

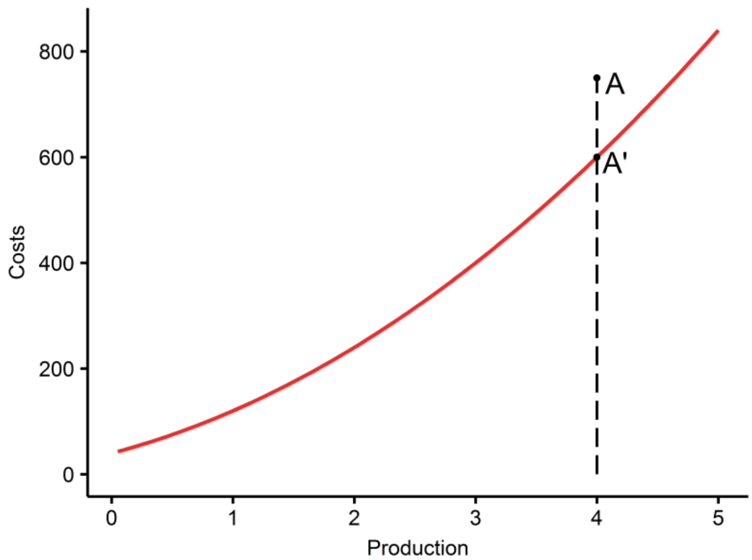
Determining the outcome of equation 2-2 is of course the choice of weights p_m and w_n. In general the weights for the resources are not much of a problem, since they can be measured by resource prices. More troublesome are the weights corresponding to the services provided or produced in public sector settings since rarely do market prices apply. Interesting methods are available to overcome this problem. A large part of this book can be directly related to this issue.

2.3 Efficiency

As seen in the discussion of productivity indices, there is a temporal aspect, namely changes over time. Within productivity, there is efficiency and technical change. Efficiency is a concept derived from productivity. It refers to the static part of differences in productivity and primarily focuses on the part of productivity that can directly be linked to management, including how to allocate resources optimally. Variations in productivity may also occur due to technological changes in time or because of different techniques, or due to environmental changes either natural or manmade (think of the influence of climate differences among countries). When calculating the relative efficiency, we have to correct differences as best we can. It is not uncommon to calculate the efficiency at an equal scale of output, or simply by comparing firms of the same size.

This concept is illustrated in Figure 2-3. The costs have been marked out against the output. The red line reflects the lowest costs at a certain level of output (the best practice). Point A is a random firm. It is clear that firm A could produce the identical output at lower costs, namely at point A'.

Figure 2-3 Efficiency



Overall efficiency can be decomposed into technical and allocative efficiency. Technical efficiency refers to the entire organization of the production process i.e., converting inputs into outputs. Technical inefficiency may be attributed to poor management or organization, leading to a waste of all resources.

Allocative efficiency relates to the right mix of resources. Sometimes higher qualified staff executes tasks that can be handled by less qualified staff. In this situation using highly qualified staff to perform simple tasks would be considered as allocative inefficient. By shifting these tasks to less qualified staff, there will be resources available to increase production. The allocative efficiency/inefficiency may also occur on the production side. By altering the output set to more profitable services, the profit or the results would be improved.

Because many public services are not directly sold, prices cannot be used to ascertain the optimal allocation of resources as in the private market. Therefore, public or non-profit organizations will increasingly be constrained by budgets, and understanding the theoretical link between technical and allocative efficiency will lead to more cost-conscious decisions. Referring back to figure 2-3, managers at firm A may either be using too many inputs to produce outputs; using too expensive inputs to produce outputs; or a combination of both. To summarize: technical efficiency refers to the right amount of resources (or services), while allocative efficiency refers to the right mix of resources (or services).

2.4 Effectiveness

Effectiveness is an administrative concept. It relates realized goals using certain policy instruments. In principle there is nothing wrong with using the same measurement instrument that is used for the conversion of resources into services. We illustrate effectiveness with the following example. In order to improve road safety, some highways have seen the introduction of lower maximum speed limits, mandatory seatbelts in the back of new cars, an increase in the number of average speed checks, and the phased replacement of inter-sections by roundabouts. If the authorities now collect the data on these variables over a number of years and regions (and/or countries) and analyze them in relation to the number of traffic fatalities and victims, the effect of each measure can be quantified statistically.

For an economist this may be a bit unsatisfactory, because he or she also wishes to assess the costs of the various methods. However, by adding price to the measures, this objection can be overcome and by using the same measurement instrument effectiveness is measured. The difference between effectiveness and efficiency is therefore the choice of the target variable. Choosing the right approach depends on which question the researcher wishes to answer.

2.5 The influence of the environment

The firm's environment also influences its productivity. For example, Dutch horticulturists use output techniques to grow tomatoes that differ from the ones used by their colleagues in Spain. Therefore, the productivity of Dutch and Spanish horticulturists differs considerably, because the variation in climate is relevant. In the provision of public services, the socio economic differences among users are often an explanation for variations in the outputs and outcomes produced. When users of public services are not functionally literate or educated, there will be communication problems or errors in the administration of services. These issues will lead to extra costs or efforts for the provider for which he cannot be held accountable. When it comes to supplying physical services, such as gas or water, differences may occur due either natural environmental factors (floods, hurricanes, and earthquakes) or manmade environmental factors (cheap materials, poor engineering).

A striking example is the output of drinking water from ground or surface water. Supplying both types of water demands a different production process in order to create drinking water. As such, environmental factors are of considerable influence on the productivity. Referring back to the spinach example, productivity and efficiency rely not only on the allocation of resources, but also on the publically sanctioned quality of the product – which includes clean water and healthy soil

2.6 Levels of research

This book mainly uses the firm as the level of research, focusing on the question: How many and which resources does a firm use, and what is the output in terms of supplied services or revenue? We focus on the consequences of management decisions and how the firm's results are influenced by the environment (for example, the labor market). However, there is no objection whatsoever to employing the same measurement instrument for questions at either the macro level or the department level. At the macro level, the question is how nationally available output resources can be employed as best as possible in order to maximize social welfare. At the department level, it can be about the optimal organization of the human resource management (HRM) department in order to realize a number of goals of staff policy (minimal sick leave, minimal staff turnover).

The following analytical levels can be distinguished:

- Macro: national;
- Meso: sectors (e.g. healthcare) or regions;
- Micro: firms (e.g. hospitals);
- Operational: departments within firms;
- Professional level.

2.6.1 Macro-level

Analyses at the macro level focus on the functioning of the entire economy, and on the optimal use of the available production capacity within the national welfare framework. For the public sector, the main questions are: What amount and what type of services should be available for an optimal public sector? How many resources must be spent on education, healthcare, and law enforcement in order to maximize welfare? It is also clear that in these analyses, special attention is paid to the influence of environmental factors on welfare. It is obvious that geographical and climatic circumstances can have a profound influence on a nation’s output possibilities.

Table 2-1 Major aspects of the research by Färe et al. (1994b)

Sector	OECD countries
Type of model	Malmquist
Data	Panel of seventeen OECD countries 1979-1988
Production	GDP
Quality	Not included
Resources	Labor and capital
Environmental factors	Not included
Efficiency factors	Not included
Economies of scale/scope	Small-scale changes of between 0.98 – 1.01
Technological development	Based on constant returns to scale technology, US total factor productivity change was 0.85 as compared to a mean of 0.70
Efficiency scores	Japan had the greatest relative efficiency change of 1.12 compared to the mean efficiency change of 0.99

An example at the macro level from the literature comes from the work of Färe et al. (1994) who assessed the technical progress and efficiency change in industrialized countries between 1979-1988. In this research, the total factor productivity of seventeen OECD countries was assessed, using the Malmquist approach and thereby controlling for inefficiency (unlike other index approaches such as the Törnqvist index) in how the countries in the sample converted capital stock and employment into gross domestic product. The findings revealed that the US led the sample countries in technical change (innovations), whereas Japan demonstrated the greatest increases in efficiency.

Another example of applying total factor productivity in a macro context comes from Khan (2005) who studied the total factor productivity in Pakistan and included other finessed measures such as human capital development (education) and economic openness (imports/exports). Given more detailed data sets, environmental factors such as the health status of the population, environmental degradation, and political freedom can also be included as outputs in measuring not only economic well-being but also social well-being.

Table 2-2 Major aspects of the research by Khan (2005)

Sector	Pakistan – macro-economy
Type of model	Growth Accounting Framework
Data	Multiple regression
Production	1960-2003 Pakistan macro determinants of GDP
Quality	Yes
Resources	Labor, capital, investment/output, Education expenditure
Environmental factors	Not included
Efficiency factors	
Economies of scale/scope	Inflation, openness to trade, investment, budget deficit, foreign direct investment, employment
Technological development	Not included
Efficiency scores	Private credit, domestic investment, government consumption, foreign direct investment, and employment all had positive effects on GDP growth.

2.6.2 Meso-level: chains and networks

At a meso-level, the structure and coherence in a sector are centralized. The most striking example is healthcare. In healthcare, different services can be regarded as complements or as substitutes. Complements are, for example, the care and treatment of patients in a home situation, in addition to formal medical treatment at a clinic. General practitioner (GP), nurse, and home- care form an entity and are therefore complementary. Substitution among services arises when one service is provided rather than another, such as a stay in a nursing or elderly home versus care provided in the patient's home. The productivity of healthcare as a whole is served by organizations and structures that include allocations of the various services for patients and clients, which is the summation of all the meso-levels, included in producing the service. Using effective substitution, healthcare must be organized in such a way that a GP can adequately treat a patient so that this patient does not end up in an expensive hospital bed or be referred to a costly specialist.

International comparisons or analyses of developments in time are fitting methods to carry out analyses at the meso level. Examples are Greene (2004) on healthcare and Afonso & St. Aubyn (2006) on education. Analyses at the meso level mostly focus on the institutional environment. Various finance systems, different forms of capacity planning, diverse ownerships, varied forms of supervision, and legal quality demands can all influence the allocation of resources (see e.g. Blank & Valdmanis, 2008).

The efficiency of chain nursing homes versus independent nursing homes was the focus of the study by Fizel & Nunnikhoven (1993). These authors sought to establish whether sharing services and reallocating patients to empty beds within the chain system led to increased efficiencies. They found that chain nursing homes were, on average, 10.3% more efficient than independent nursing homes. The reason for this better performance was attributed to scale efficiencies. In another study of long term care facilities, Dervaux et al. (2006) found that these facilities would be more efficient if patients were admitted to appropriate level of services according to their individual health care needs.

Table 2-3 Major aspects of the research by Fizel and Nunnikhoven (1993)

Sector	Nursing home chains
Type of model	Input based DEA
Data	1987 annual year-end costs Department of Long Term Care Settlements, Michigan, US
Production	Number of patients/residents
Quality	Percent skilled beds
Resources	Aides, licensed practical nurses (LPNs), registered nurses
Environmental factors	Not included
Efficiency factors	Chain membership, ownership
Economies of scale/scope	Not included
Technological development	Not included
Efficiency scores	Overall efficiency = 0.66; individual nursing homes = 0.62; chain nursing homes = 0.71

2.6.3 Micro level: firms

The micro level focuses on firms. It is mostly about an economic-judicial unit within which there is centralized decision-making, such as a board of directors that makes the major decisions about management and strategy. In other words, the corporate process regarding how resources are converted into actual services or products is centralized. Economic considerations, when management strives for a certain goal and must take numerous technical and institutional constraints into account are the starting point of the analyses. In Florida, there are public health centers for each county. In a study of capacity and change required by the Affordable Care Act of 2010, Valdmanis et al. (2015) found that public health clinics could provide more direct patient care with existing capacity, but there was not the required demand to efficiently use the public clinics' services which is only one part of the public health centers' obligation.

A common example of efficiency analysis at the firm level is the banking sector, including banks and bank branches. For the comparison of banks, the regulation of financial markets in different countries is often centralized (Fitzpatrick & McQuinn, 2008; Lensink et al., 2008). When comparing bank branches, the question is often how the various branches should be directed from the main office, and what influence the environment has on the results of the branches. But, as mentioned above, the banks play a major role in safe-guarding the welfare of society.

2.6.4 Operational level: departments

The operational level relates to the departments, sections, and locations of a firm. All HRM, personnel, and salary departments can be compared, as the services they supply are rather homogenous. Salary payments, changing staff data as a result of hiring and firing, changes of positions, or organizing courses for staff members are examples of these services. The productivity of a department is then measured against the supply of intermediary services. The supply of the end service of the firm concerned plays no part in determining operational levels.

What is important in ascertaining how well departments within a firm perform, is the level of cooperation and coordination among departments in meeting (1) the objective function of the department coupled with (2) the objective function of the firm in general (Tjosvold, 1988). Thus, productivity analysis needs to be teamed with the organizational behavior literature in order to fully understand the dynamics underlying efficient practices.

2.6.5 Professional level

The most micro level is that of the professional. The achievements of doctors, teachers, and judges can be compared using individual data. Their individual talents, experience and capacities in their profession are of vital importance to the outcome of the entire process. Didactic and pedagogic methods may differ considerably among teachers and are often vital to the students' performance. An example is specialists' experience certain complicated procedures. For some specific treatments, the Dutch Inspectorate for Healthcare applies a minimum standard for the number of treatments a specialist must perform each year in order to be able to perform the treatment at all. In the US, pay-for-performance would require that physicians and hospitals be reimbursed only for outcomes excluding any extra costs due to poor quality of care. One such change is the Hospital Readmission Reduction program § 3025 of the ACA. This section was written into the ACA in order to provide quality incentives to hospitals for reducing the costly and largely preventable re-admissions of patients (Kocher & Adashi, 2011). Penalties imposed on hospitals experiencing re-admission rates were 1% of hospital aggregate Medicare payment in 2013, increasing to 2% in 2014 and 3% in 2015 (Nuckols, 2015).

Re-admission rates are used for assessing hospital quality particularly if they are related to infections of other complications which have been shown to exist with lower levels of hospital quality (Davies et al., 2013). The 30 day re-admission rate is the most commonly used as this rate is representative as arising when preventable measures are not taken during the patient's initial hospital stay (Benbassat & Taragin, 2000; Davies et al., 2013; Herrin et al., 2015; Weiss et al., 2011). Reasons for lower quality include nursing (RN) staffing and providing effective patient education on the importance of medication compliance and follow-up care with their primary care provider (Weiss et al., 2011).

Table 2-4 Major aspects of the research by Chilingerian (1995)

Sector	Physicians' practices
Type of model	Output based DEA
Data	Physician data from a hospital in Boston
Production	Low and high severity of patients
Quality	Not included
Resources	Total length of stay; total ancillary services
Environmental factors	Not included
Efficiency factors	Physician affiliation with HMO; physician Age; size of case load; surgeons versus. internists
Economies of scale/scope	Average scale efficiency ranged from 88-94%
Technological development	Not included
Efficiency scores	An average 80-95% efficient depending on the model;

Chilingerian (1995) analyzed the productivity and efficient practices of physicians. He found that were all physicians to practice and use resources in the way that the most productive physicians do, overall health-care costs could be greatly reduced. This work is exemplary of how individuals can be tracked in terms of productivity research, and of how lower costs can be realized, if individual behavior can be changed. It should also be noted that treatment practices included in the Affordable Care Act (the US health-care reform program) include measures of physician efficiency and quality. The approaches proposed by Chilingerian can be used by analysts who are focusing on these types of research issues. For a more recent treatise on physician productivity using DEA, see Sherman & Zhu (2006:215-241).

3 Microeconomic theory

3.1 Introduction

As stated earlier, this book is an introduction to the use of productivity analysis for public (or social) sector decision-making. Although not meant as an economics or public finance text, reviewing the basic tenets of microeconomics is relevant. Therefore, in this chapter we focus on reviewing the economic theory of the firm, which is also known as producer's behavior. The question how a firm should convert resources into products and services is of central interest. This concerns purely technical matters as well as the economic behavior of the firm and the (institutional) environment. The chapter also deals with the way these theoretical insights can be turned into an economic model that makes it possible to conduct quantitative empirical research on productivity in the public sector.

3.2 Economic theory of the firm

3.2.1 A few definitions

Before proceeding to the theory of the firm, we present the definitions of the family of costs.

Fixed costs are the costs that do NOT change with the amount produced by an organization. Examples of fixed costs include buildings, rent, and other non-negotiable costs for a period of time.

Average fixed costs are derived by dividing fixed costs by the quantity an organization produces.

Variable costs are the costs associated with the production of the quantity made. These costs typically include labor, equipment, and utilities.

Average variable costs is simply variable costs divided by the quantity an organization produces. Note that this concept only applies in the case a firm produces one type of output. In case of multiple outputs other methods need to be followed.

Total costs equal fixed costs plus variable costs; average total costs equals total costs divided by quantity produced.

Perhaps the most relevant for decision makers are *marginal costs* which is defined as the change in costs by producing one more unit of output. The importance of marginal costs is when these costs begin to rise over average variable costs, and price is equal to marginal costs; the decision maker will stay open to cover fixed costs. If price falls below average variable costs, the decision maker must shut down immediately.

These costs are typically applied to for-profit firms. However, costs are influenced by productivity and efficiency which may be subject to budget constraints or re-assessment by decision makers whether these services are needed at current levels.

3.2.2 The production function and the technology set

A production function explains the relation between the amount of resources employed and the amount of goods or services produced. In Figure 3-1 we show this process using the example of the relation between the number of square meters of tilled public park (services/output produced) in a year, and the number of employees, expressed in full-time equivalents (FTEs) (amount of resources/inputs).

Figure 3-1 Relation between square meters of green area and employees of park-keeping service

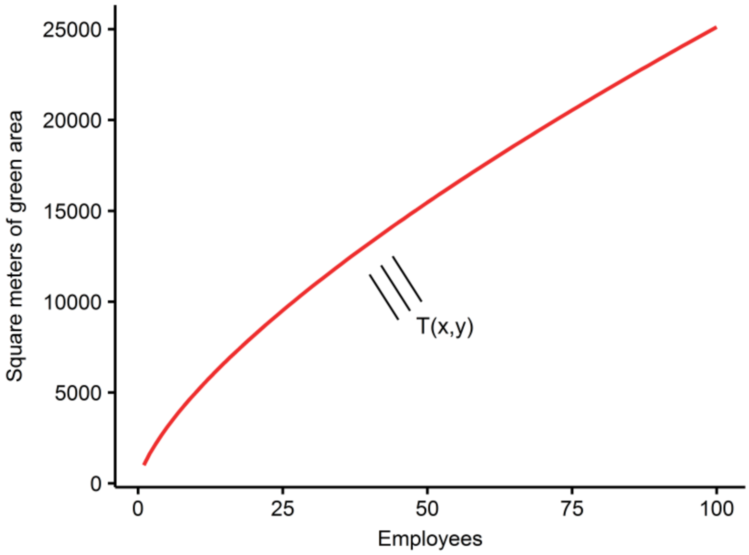


Figure 3-1 shows that 10 employees till about 5,000 square meters, whereas 100 employees till a little over 25,000 square meters. The increase in services produced, decreases when the number of employees increases; this is referred to as diminishing returns to scale (scale effects) and is described in subsection 3.2.3. The red line indicates the maximum feasible output at a certain effort expended by the employees. Producing less is, of course, also possible. In fact, all points below the red line are feasible, but are considered inefficient. This set of input and output combinations is known as the technology set T(x,y).

Naturally, the park-keeping service uses not only human labor but also tools and therefore this relationship requires the use of an isoquant, which is a line that reflects all possible input combinations that lead to a similar output level. In Figure 3-2 we illustrate such an isoquant using the combination of capital depreciation on the Y-axis and labor on the X-axis. Because tools and transport can often be used for several years, the employment of these resources equals their depreciation.

Figure 3-2 Feasible and minimal combinations of capital depreciation and employees at given level of square meters tilled park (Input set and isoquant)

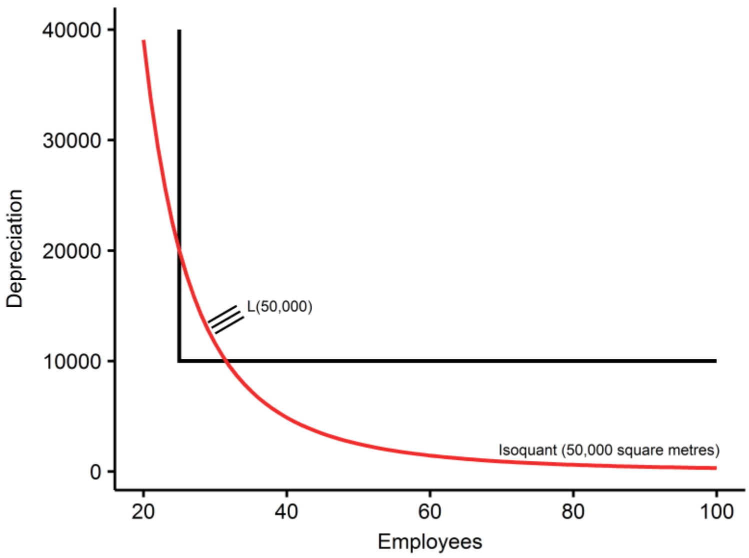
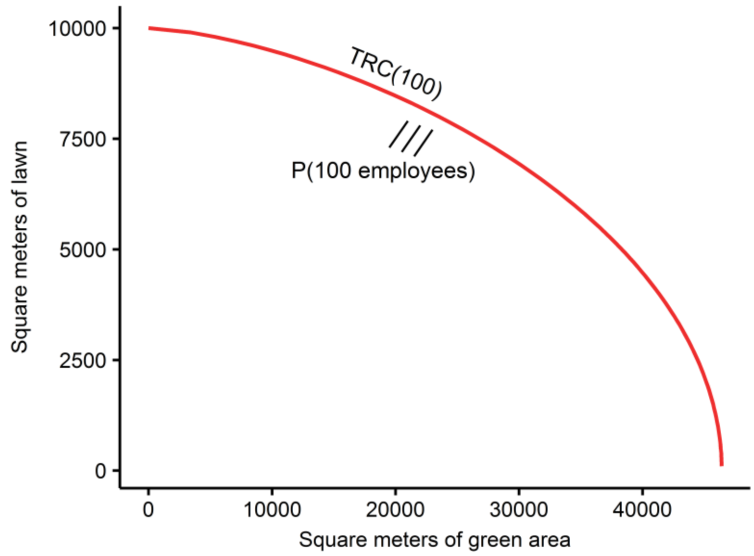


Figure 3-2 shows all combinations of numbers of employees and depreciations that in turn result in 50,000 square meters of tilled public park. The nature of the output and the available technology determine the isoquant's form. Because of the curvature presented by the smooth figure, there are possibilities to exchange labor and capital, which is referred to as substitution. However, there are production processes where substitution is not possible. In that case the figure is represented by the rectangular figure. But in all cases, observations located inside the isoquant, (i.e. to the right of the red line), use more employees and/or tools to generate an output level of

50,000 square meters, which is why they are included in the input set L(50,000), but are deemed inefficient.

Suppose the park-keeping service must also provide additional services so that besides tilling it must also, regularly mow the various municipal lawns, we then speak of multiple production or *multiple outputs*. In that case, the park-keeping service must decide how to allocate the labor and capital used to till and to mow the lawns. Analogous to the previous example, all combinations of services can be represented in a graph, with which a fixed input set is assumed. Figure 3-3 depicts the output set and the transformation curve which is analogous to the production possibility curve..

Figure 3-3 The output set and the transformation curve (TRC)



In contrast to Figure 3-2, where inputs are on each of the axes, in Figure 3-3 we show the example of a park-keeping service with 100 employees producing the two outputs on the axes. The red line reflects all maximum feasible combinations of square meters of tilled public park and mowed lawn. Note that the production possibilities curve depicts the combination of outputs that can be produced holding inputs fixed. In principle, the combinations below the transformation curve are also possible; they are included in the output set P (100) but are considered inefficient, and can reflect wasted resources or non-productive worker time.

How these curves in 2-dimensional space in multi-dimensional space (it is possible to calculating in more dimensions geometric forms called polyhedrons) are constructed depends on the production function, which demonstrates how specifically inputs are transformed into outputs. A very well-known mathematical presentation is the Cobb-Douglas (CD) production function (Cobb & Douglas, 1928). It is defined as:

$$y = a_0 x_1^{\alpha} x_2^{\beta} x_3^{\gamma} \qquad (3\text{-}1)$$

In which y is the output and the various x' s represent the resources used to produce this output, such as the number of public-park employees (labor), the used fertilizer and the wear and tear of the lawn mowers (capital). The other letters a_0, α, β, and γ are called the parameters of the function. A CD function was used when creating Figure 3-1. When we, for example, fill in 1,000 and 0.7 for a_0 and α and calculate the corresponding y value for each value of the number of employees [1, ... , 100], Figure 3-1 can easily be reproduced.

In the public sector, organizations frequently produce more than one output and we have alluded to this previously in discussing multiple outputs. We can expand the CD production function as follows:

$$1 = a_0 y_1^{\delta} y_2^{\varepsilon} x_1^{\alpha} x_2^{\beta} x_3^{\gamma} \qquad (3\text{-}2)$$

This specifies that each combination of y's and x's must result in a value of exactly 1 because the value of 1 is assigned to an efficient firm. Equation (3-2) can be rewritten with for example y_2 on the left hand side. In that case, y_2 depends not only on the employment of resources, but also on output y_1. If y_1 is substantial, fewer resources are available to produce y_2 , a situation that often arises in the competition for resources between private and public producers. We provide an example of this in Table 3-1 using the following parameters: a_0 = 1.58; δ = 0.6; ε = 0.4; α = -0.2; β = -0.4; γ = -0.6. These are made up numbers used for illustrative purposes and not derived.

In Table 3-1, we show that there are many plausible options for producing the outputs. Firm 1 is characterized by using a relatively large amount of resource 3 (x_3 = 82) and an almost equal output of product 1 (y_1 = 63) and product 2 (y_2 = 67). Firm 3 also uses relatively more of resource 3 (x_3 = 79), but produces a great deal of product 2 (y_2 = 191). As far as output goes, firm 3 resembles firm 9, but firm 9 uses a relatively large amount of resource 2 (x_2 = 68) and 3 (x_3 = 80) and very little of resource 1 (x_1 = 23).

Table 3-1 Example of multiple inputs and multiple outputs production

Firm	x_1	x_2	x_3	y_1	y_2
1	21	31	82	63	67
2	34	23	35	43	31
3	41	42	79	46	191
4	29	45	64	76	59
5	54	64	35	85	39
6	35	10	53	64	14
7	46	23	45	52	40
8	52	52	26	66	29
9	23	68	80	58	167
10	10	42	19	60	7

Technologies may vary considerably among sectors and over time, with some technologies being more labor intensive and others being more capital intensive. Hence, the effect of the employment of resources will vary substantially. Many mathematical specifications have been applied in order to map all these technologies it would be going too far off the subject to discuss them extensively. Common examples are the Leontieff function, the constant elasticity substitution (CES) function, the quadratic function, the Fourier function, and the translog function (the most popular one). This specification will be further explained in the Annex of chapter 3. Readers who are more interested in the mathematical details are referred to the standard books mentioned in Chapter 1(Fried et al., 2008).

Recall that, in Equation (3-2), we illustrated the relation between resources and output for efficient firms. Such visualization was used by Shephard (1953) in order to show all possible input and output combinations, as well the ones that can be labeled inefficient. For this, he introduced the notion of the output distance. Suppose that firm 1 produces 10% less (respectively 56.7 instead of 63, and 60.3 instead of 67 units), then Equation (3-1) will result in a value of 0.90 instead of 1. This is the "production distance" or "output distance" of the firm to the efficient frontier. The farther away from the frontier, the greater the distance function i.e., inefficiency. The output distance reflects the smallest number by which the number of services/goods can be divided without requiring more resources.

Given this formal definition, we provide the following example to illustrate the concept. This firm performs at 90% of the distance to the efficient firm. Distance, then, is a relative concept and is presented in Figure 3-4. When the output is reduced by half and the employment of resources remains unchanged, the equation will result in a value of 0.50. This relationship is demonstrated in Equation 3-3, which is as follows:

$$D_o = a_0 y_1^{\delta} y_2^{\varepsilon} x_1^{\alpha} x_2^{\beta} x_3^{\gamma} \qquad (3\text{-}3)$$

(a_0 = 1.58; α = –0.2; β = –0.4; γ = –0.6; δ = 0.6; ε = 0.4)

In which:

D_0 = output distance.

The expression in Equation (3-3) reflects an output orientation (maximizing services at given resources), because the sum of the parameters of the products y_1 and y_2 δ (=0.4) and ε (=0.6) equals 1. This requirement is defined as homogeneity of the degree 1 in y_1 and y_2, meaning that when both services are increased by 10%, the output distance will also be increased by 10%, which is precisely in line with how output distance is defined. These and other requirements will be discussed in detail in Section 3.8.

When all resources in a firm are doubled, then applying Equation (3-3) generates a value of 0.44 and not a value of 0.50. The reason why will become clear later on in Section 3.8.

Besides the output distance, there is also the notion of input distance, which is given as minimizing inputs at given outputs. An input distance function is formulated in such a way that doubling the resources will lead to a value of exactly 2. Equation (3-3) may also be formulated from a resources perspective, given as:

$$D_i = a'_0 \, y_1^{\delta'} y_2^{\varepsilon'} x_1^{\alpha'} x_2^{\beta'} x_3^{\gamma'} \qquad (3\text{-}4)$$

(a_0' = 1/1.46; α' = 1/6; β′ = 1/3; γ′ = 1/2; δ′ = –1/2; ε′ = –1/3)

In which:

D_i = input distance.

Now, the parameters α', β', and γ' add up to 1. Note that filling in the original data of firm 1 then leads to the value 1. Doubling the resources here will lead to an input distance of 2. Cutting the output of firm 1 by half at an equal input of resources will result in a value of 1.78 (take y_1=31.5; y_2=33.5 and calculate equation (3-4)). The output distance always indicates a value less than or equal to 1 (and greater than 0) and the input distance always indicates a value greater than or equal to one. We come back to this feature later on as well.

Input oriented distance functions also meet the homogeneity (of the degree one) requirement, but now in the variables x_1, x_2 and x_3.

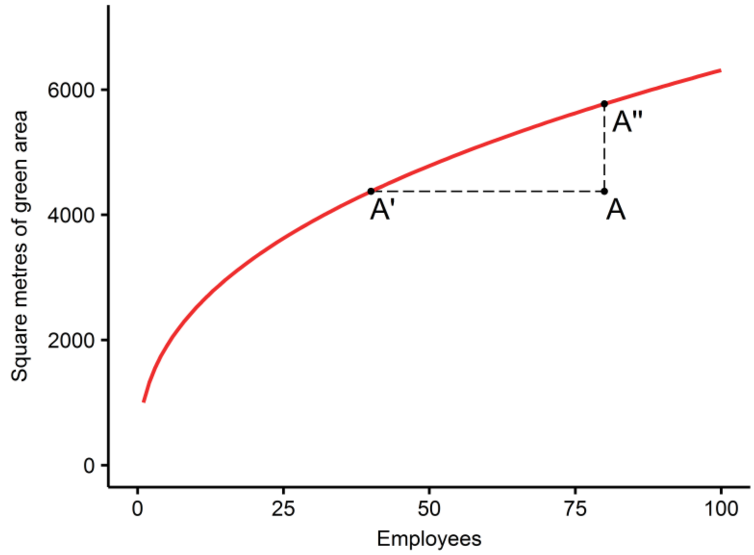
3.2.3 Technical efficiency

Both distance measurements play a part in determining the technical efficiency of a firm. We speak of technical efficiency when it is not possible to increase the output of a product without decreasing the production of at least one of the other products, or increasing the employment of at least one resource. The formulation can also be reversed. We speak of technical efficiency when it is not possible to decrease the employment of a resource without increasing at least one of the other resources, or decreasing the output of at least one of the products. This is known as the Koopmans definition of efficiency (Koopmans, 1951).

The Koopmans definition of efficiency is a very formal one. In lieu of using the Koopman's definition, authors of many empirical works utilize the Debreu-Farrell definition (Debreu, 1951; Farrell, 1957), which offers a standard for the technical efficiency. In the Debreu-Farrell context, the technical efficiency equals 1 minus the proportional decrease of all resources without decreasing the output. (It can also be applied to maximizing output without increasing inputs.) The difference between the Koopmans definition and that of Debreu-Farrell, is that the latter definition adds the notions "proportional" and "all". According to the Koopmans definition, there only needs to be a decrease in one resource or increase in one product, whereas according to the Debreu-Farrell definition there must be a proportional decrease or increase in all resources or products. This proportional reduction in all resources is important as it keeps the technology the firm uses intact. For example, if a firm uses six employees and six tractors but it is inefficient by 33%; the firm needs to reduce employees and tractors both to four to maintain a technology of one employee per tractor.

In Figure 3-4 we elaborate on the Debreu-Farrell definition.

Figure 3-4 Input distance, output distance and technical efficiency



The red line in Figure 3-4 reflects the production function.. Point A represents the combination of the number of employees and the generated output of firm A. It is clear that firm A can perform better. For the given employment of staff, an output of firm A″ is feasible. The distance between A and A″ is a measure for A's output oriented inefficiency. When we reverse the efficiency notion, we see that firm A can generate the same output level with fewer employees, namely the number of employees of firm A′. Therefore, the distance AA′ can also be regarded as a measure for A's input-oriented inefficiency. It need not be said that both distance measurements can differ. From this we can deduce an important notion. The efficiency to be calculated is partly determined by the direction along which it is measured. This is known as the direction or orientation. (For AA″ we speak of an output orientation, for AA′ of an input orientation.) The calculated efficiency is determined by the orientation or the reference point the researcher chooses. The question which orientation is the right one depends on the context. A typical example of output orientation is when a school receives a fixed budget or a fixed number of staff members. The only margin the school has is to employ the available resources in such a way that a maximum education result can be realized. Alternatively, an inefficiency score according to the input orientation can be applied if the government does not fix the number of staff members, but the amount of output(s) produced is fixed.

The distance function comes in handy when the one wants to calculate the technical efficiency. That is, the calculated output distance D_0 equals the ratio between the y-value of A and A'':

$$D_o = \frac{y_A}{y_{A''}} \qquad (3\text{-}5)$$

This means that D_o always assumes a value between 0 and 1 with the value of 1 equaling efficiency.

The input distance D_i is defined as the ratio between the x-value of A and A′:

$$D_i = \frac{x_A}{x_{A'}} \tag{3-6}$$

Note that both distance measures take the ratio between the actual value and the frontier or best practice value. In case of the output distance this leads to a number smaller or equal to one, whereas in the case of the input distance it leads to a number greater or equal to one. So please note that distance measurements have a different value range. The input distance intuitively seems to be the right measure, because a greater distance in Figure 3-3 also matches the greater value of the distance measurement whereas, the output distance is counter-intuitive, because a greater distance matches a smaller distance measurement. One benefit of the output distance function is that it exactly equals the output-oriented technical efficiency. For the input-oriented technical efficiency, we need to utilize the inverse of the input distance ($1/D_i$), and the calculated inefficiency scores will be less than 1. In the previously calculated example, it appears that when the resources are cut by half (at an equal output), the input-oriented technical efficiency also results in a value of 0.5 (=1/2), whereas the output-oriented variant results in a value of 0.44.

So, when all products increase by 10% without any extra resources, the output distance should also increase by 10%. In equation 3-3 this shows because the sum of δ (= 0.4) and ε (= 0.6) equals 1.

Using the distance function approach, Dervaux et al. (2004) compared the technologies of French vs. US hospitals, and identified the sources of the inefficiencies. Because of the attributes of the distance function, inefficiency scores can be added to arrive at a grand total of inefficiency (see Table 3-2).

Table 3-2 Major aspects of the research by Dervaux et al. (2004)

Sector	French and US hospitals
Type of model	Distance function
Data	1080 French hospitals (Statistique Annuelle des Etablissements); 903 US hospitals (American Hospital Association data)
Production	Admissions, births, inpatient surgeries, outpatient surgeries, emergency room visits, outpatient visits, medical interns and residents
Quality	Not included
Resources	Beds, physicians, nurses, other personnel (measured in full-time equivalent)
Environmental factors	Ownership, country
Efficiency factors	Not included
Economies of scale/scope	9.5% of total French hospital inefficiency attributed to scale inefficiency; 6.3% of total US hospital inefficiency attributed to scale inefficiency
Technological development	Not included
Efficiency scores	French hospitals: overall inefficiency: 19.8% (congestion: 3.2%, technical, 7.1%) US hospitals: overall inefficiency: 23.7% (congestion: 3.4%, technical 14.0%)

3.2.4 Scale and diversification effects

Returns to scale refer to the relation between resources and scale (range) of output. They indicate by which factor the output changes when there is a proportional change in all resources. In other words, when the output changes by the same factor as the resources, we speak of constant returns to scale. When the change is less than proportional, we speak of decreasing returns to scale. Increasing returns to scale indicate that the output grows faster than the increased employment of resources. The often positive returns to scale in smaller firms can be explained by the increasing possibilities to distribute labor and to make more efficient use of buildings and machines.

Decreasing returns to scale in larger firms may be due to the growing bureaucracy or to distractions among many more employees. Between these two extremes, we often speak of an optimal scale corresponding with a maximum benefit from the distribution of labor without the negative influences of bureaucracy.

A very common way to express the relationship between economic variables is by using elasticities. Elasticities represent the percentage change in one variable as the result of a percentage change in another variable. The elasticity of scale measures returns to scale as the percentage change in production as a consequence of a change in all resources by 1%. A value smaller than 1 defines decreasing returns to scale, a value equal to 1 with constant returns to scale, and a value greater than 1 defines increasing returns to scale. The elasticity of scale can be deduced from the production function by determining how it changes when each of the resources changes marginally.

$$\boldsymbol{EOS} = \frac{\%\ \boldsymbol{change\ output}}{\%\ \boldsymbol{change\ all\ inputs}} = \frac{\frac{\boldsymbol{dy}}{\boldsymbol{y}}}{\frac{\boldsymbol{dx_1}}{\boldsymbol{x_1}}} + \frac{\frac{\boldsymbol{dy}}{\boldsymbol{y}}}{\frac{\boldsymbol{dx_2}}{\boldsymbol{x_2}}} + \frac{\frac{\boldsymbol{dy}}{\boldsymbol{y}}}{\frac{\boldsymbol{dx_3}}{\boldsymbol{x_3}}} =$$

$$\frac{\mathbf{x_1}}{\mathbf{y}}\frac{\mathbf{dy}}{\mathbf{dx_1}} + \frac{\mathbf{x_2}}{\mathbf{y}}\frac{\mathbf{dy}}{\mathbf{dx_2}} + \frac{\mathbf{x_3}}{\mathbf{y}}\frac{\mathbf{dy}}{\mathbf{dx_3}} \qquad (3\text{-}7)$$

Suppose we have a production function:

$$\boldsymbol{y = 1.58 \cdot x_1^{0.2} x_2^{0.4} x_3^{0.6}} \qquad (3\text{-}8)$$

We elaborate the first term:

$$\frac{x_1}{y}\frac{dy}{dx_1} = \frac{x_1}{y}[0.2 \cdot 1.58 \cdot x_1^{-0.8} x_2^{0.4} x_3^{0.6}] =$$

$$\frac{x_1}{y}\left[0.2 \cdot 1.58 \frac{x_1^{0.2}}{x_1} x_2^{0.4} x_3^{0.6}\right] =$$

$$0.2 \cdot \frac{x_1}{y}\frac{y}{x_1} = 0.2$$

If we elaborate each term on the right hand side then:

$EOS = 1.2$

If we start from the output distance function – which we need in case of multiple outputs – then the expression with respect to the economies of scale is:

$$\varepsilon_s = -\left[x_1 \frac{\partial D_o}{\partial x_1} + x_2 \frac{\partial D_o}{\partial x_2} + \cdots\right] \quad (3\text{-}9)$$

In which:

ε_s = elasticity of scale.

The term $\frac{\partial D_o}{\partial x_n}$ reflects the (marginal) change in the output distance when the use of resource n increases by a small amount.

Please note that when the employment of resources increases (e.g., more employees in Figure 3-4), the distance in Figure 3-4 increases and the output distance therefore decreases. This explains the negative sign in equation (3-9).

An approach via the input orientation is also possible, resulting in the following equation:

$$\boldsymbol{\varepsilon_s} = \frac{-1}{y_1 \frac{\partial D_i}{\partial y_1} + y_2 \frac{\partial D_i}{\partial y_2} + \cdots} \quad (3\text{-}10)$$

Please note that an increase in output is accompanied by a lower value of the input distance. Therefore the denominator in Equation 3-10 is always negative; hence, the minus sign in the numerator.

When we apply these formulas to the example in Table 3-1, it appears that the elasticity of scale equals 1.2. This means that a 1% growth in the employment of resources leads to a 1.2% growth in output, indicating increasing returns to scale.

Diversification relates to the influence of the mix of services on productivity. The central question here is whether, at a given quantity of resources, the productivity can improve by specialization in some services or rather by producing services jointly. Diversification effects often occur because of a partial or joint use of resources (Lovell, 2000), also known as economies of scope.. One of the well-known examples is sheep herding. The sheep can be held for slaughter as well as for the production of wool. Meat and wool production mix well in this case (economies of scope). From a productivity point of view, the specialized separate production of wool and meat is not optimal. Conversely, the combination of a welding firm and fireworks storage would require a large number of extra safety measures that would negatively influence the productivity. In this example specialization is more obvious (diseconomies of scope).

A way of determining economies of scope is via cost complementariness. Cost complementarities arise when an increase in the production of service 1 leads to a decrease in the marginal costs of service 2. The combined output of both goods then leads to cost benefits. Conversely, when the marginal costs of service 2 increase at an extra output of service 1, we speak of diseconomies of scope (or a negative diversification effect). In that situation, it is feasible to produce the goods separately (specialization). In formula:

$$\boldsymbol{CostCompl} = \frac{\Delta\left(\frac{\Delta C}{\Delta y_2}\right)}{\Delta y_1} \tag{3-11}$$

When the previous expression is less than 0 (zero), we speak of economies of scope or positive diversification effects. Note the confusing formulation in the definition: a negative outcome of (3-9) matches with positive effect!

Coelli & Fleming (2004) suggested an alternative approach by starting on the technical side by using an output distance function, which is more in line with the original definition of scale effects. This approach to diversification has not been used very often in research and we leave it here further unspoken.

An interesting alternative is the all-or-nothing option, when the total input of resources of fully specialized firms is compared to that of firms that have produced the same amount of services jointly (see Baumol et al., 1988). There are economies of scope when the costs of joint production of two different products are lower than the costs of the separate production of these two products. The disadvantage of this approach is that economies of scope and economies of scale intertwine. It is obvious that a firm that produces both products often does so on a larger scale than a firm that produces only one product. The combination in itself is not relevant in such a situation.

For an extensive discussion of the estimation of economies of scope, we refer to an excellent overview presented by Triebs et al. (2012).

3.2.5 Technical, social and institutional changes

Technical, social, and institutional changes influence production methods. We can think of new machines that simplify the work, such as computerization for measuring vital signs and electronic alarm systems in the care of the elderly. Thanks to such an increase in technological knowledge, various resources will be put to better use. A different look at the quality of service can also be of influence on the output structure. Social preferences play a part in this. Legislation also frequently changes. Numerous measures aimed at creating a safe and healthy environment go hand in hand with other production methods and often involve extra costs. Such shifting is called technical or autonomous change and happens over time. This technical change differs from technical efficiency. The technical efficiency reflects the distance to the best practice, whereas the technical change shows a shift of this best practice in time. Readers are again referred to the work by Färe et al. (1994), which we cited in section 2.6.1 on changes in the OECD countries' total factor productivity over time. Another reason for technological change is the increased use of automation to replace manufacturing jobs. Current politics suggest that the loss of US manufacturing jobs is due to trade – not so, it is due more to technology changes.

3.3 Economic behavior

The previous section was entirely focused on the technical relation between inputs and outputs, and completely ignored such economic questions as: How many products must a firm supply, and which resources must be employed to achieve this? Answering these questions is linked to the economic objectives of the firm. Firms in the market sector are mostly driven by realizing profits, or selling inventory, or gaining market share. In the public sector, firms mostly focus on the public interest and financial aspect preconditions rather than goals. It is clear that different goals may lead to different outcomes. We discuss these goals below.

3.3.1 Output maximization

Output maximization relates to the concepts presented in section 3.2. Maximizing the output is accomplished by finding out by what factor the output of all products may grow maximally without adding any extra resources. This is the previously discussed definition of output-oriented efficiency and is applicable when the employment of separate resources has been determined beforehand that is set by a budget. In some situations, the staff composition as well as the number of square meters of building surface are predetermined. It should be noted that output maximization is a necessary but not sufficient, condition for revenue maximization.

3.3.2 Input minimization

Input minimization is also directly linked to the technical efficiency approach of Section 3.2, which determines the factor by which all individual resources can be reduced without affecting the output of products. An example is reducing 'scrap' of resources. Think, for example, about decent working conditions. Working with good and safe equipment and decent personal protection equipment may lead to fewer defective machines and fewer sick employees. Unsafe working conditions lead to public intervention such as the Occupation, Safety, and Health Act (OSHA). Similar to the case of output maximization, input-oriented efficiency is a necessary, but not sufficient condition for cost minimization.

3.3.3 Revenue maximization

Revenue maximization focuses on striving for maximum return at a given employment of resources. The use of resources is fixed and the firm determines the services mix. A condition is that prices are known for the different products. An interesting example is the study by Van Tulder (1994), who researched police efficiency. Before the reorganization of the Dutch police force in 1994, the government managed the employment of resources, and determined the number of executive and non-executive staff, as well as the use of material resources. With the determined resources, according to Van Tulder (1994), the police tried to realize as much "output" as possible, with which the different products were weighted with prices linked to fines, jail time, and/or damages, all valued in terms of money.

Since the reorganization, local forces have had more discretion about how to distribute the lump sum over the various resources. This approach of allocating a fixed budget is referred to as indirect revenue maximization. At an available cost budget, a firm will try to maximize its revenues or output since police are not typically profit maximizers. It decides not only the optimal level of the services, but also upon the optimal mix of the services.

3.3.4 Cost minimization

Cost minimization implies that the firm's goal is to minimize costs while producing a given level of services. The level and mix of the services produced are fixed, but the input of resources is not. Prices of the resources play an important role. As the price of a resource increases, the firm will try to use less of that resource. It will compensate for an increase in wages by employing labor-saving capital goods.

There is also a variant in which revenue is linked to a certain production of services. In the past, production agreements, which were translated into a revenue constraint, were arranged with Dutch hospitals. The total of invoiced revenues of a hospital was not allowed to exceed the annual revenue constraint.

3.3.5 Profit or disguised profit maximization

The model of profit maximization is the most common model of microeconomic theory. Based on given prices of products and resources, the firm determines how many resources it will employ and what production level that will achieve. Managers decide upon the allocation of products and resources, and on the levels of products and resources.

In the public sector, this application is not typically viable. In most cases, the firms are not allowed to distribute profits. They must consider social welfare, for which prices may not be readily evident, and other social constraints that have to be taken into account. The cost level is predetermined (through budgets), the capacity is determined by the government (employment of capital), or the production of services has been fixed. That leaves little room for firms to make decisions about the allocation of resources or products as they deem fit.

While it usually is not permitted to distribute a profit, there are ways to earn net revenue. Surpluses from the net revenue can sometimes be used to build reserves, to increase the salaries of executives and managers, or to finance improvements in equipment, corporate resources or housing. In the USA, the key legal requirements state that non-profits can earn net revenues, but those revenues cannot be distributed outside the organization to, say, stockholders. However, excess revenues can be used to expand services, hire more nurses, staff, et cetera. In all these cases, volume can be enhanced to the benefit for the community.

3.3.6 Input preference

The model of input preference starts from the idea that a firm tries to maximize the employment of resources. An example of input preference follows from bureaucratic behavior: A firm’s management will try to achieve as much prestige, status, or salary as possible by expanding its department budgets or employment of staff (Niskanen, 1968). Examples of such research are Rodriguez-Alvarez & Lovell (2004) and Blank et al. (2012). Rodriguez-Alvarez & Lovell researched the preference of management in Portuguese hospitals, while Blank et al. studied the management and the support in Dutch secondary schools.

This preference for a certain resource often leads to a suboptimal allocation of resources. By using a different mix of resources, a firm could generate better performance or the same performance at lower costs. Interestingly, it follows from the study by Blank et al. (2012), that in secondary education there is no systematic preference for management or support. Similar results can be found in Haelermans et al. (2012). Another example of input preference is the value of employment opportunities provided by public entities, particularly in rural or economically distressed areas.

3.3.7 Maximization of public values

The above-mentioned behavioral assumptions relate exclusively to the financial aspect of output, in which case output can also be expressed in terms of quality or efficiency. We can take yet another step and explicitly include the social interest of the service. Appreciation for the services may be higher than exclusively the economic value in terms of the prices paid to the producer. Moreover, public prices may differ proportionally from economic prices. This means that for example a certain combination of treated patients might not be optimal from a business economic point of view, but the number of individuals receiving services might be optimal from a social point of view. A good example is the increased emphasis on mental health with parity of physical health since lack in either or both lead to decreased productivity and well-being.

This type of models has not been developed very well yet, but research in this area is progressing. Blank (2013) has applied this idea to social job creation in the Netherlands, which meant weighting different types of placements (internal, special working agreements for individuals with physical or mental disabilities) differently, whereas the government subsidy for each placement is the same.

3.4 Constraints

Constraints refer to the degree of freedom in a firms' decision-making. In a fully free market firms are free to decide what they produce, the level of quality they supply, what price they charge the client, how they produce, and what resources they use. It is easy for free market firms to do this because all information is conveyed by the price of the good/service and consumers can judge the quality of the good/service in accordance with their tastes. In practice, firms will always have to deal with constraints resulting from, such external effects as the environment, labor conditions, contracts, etc.

In the public sector, there are additional constraints that must match certain public values, such as accessibility, quality, and affordability. Hospitals are obligated to give medical attention to those who need it, irrespective of the patients 'ability to pay or the fact that he or she might be costly to treat. In the US, federal law mandates that hospitals cannot refuse patients presenting to an emergency room, just because they may incur high cost to the hospital.

Constraints are imposed at different stages in the production process. It sometimes means that a certain output must be realized, whereas in other cases the employment of resources or the available budget has been meticulously laid down. Besides the economic behavior and the resulting optimization behavior, a number of other constraints are important, namely:

- Service constraints;
- Resource constraints;
- Budget constraints;
- Revenue constraints;
- Profit constraints.

When a public school has the legal obligation to accept all students and not just the highly talented ones, the output cannot be influenced much. In that case, the school must supply what is required. The output is fixed. But there may special requirements mandated by law such as special needs education. Here is a case for education policy makers need to use the concept of economies of scope.

Sometimes the *output or service constraints* go further and then even the intermediary output is documented. Intermediary outputs in the case of education include such products, as the number of hours of teaching.

When the resources have been documented in detail, we speak of *input or resource constraints*. Examples are documented staff composition and the material defrayment in education, indicating that the extent of the resources is fixed (lump-sum payment) and a trade-off between resources is not possible.

Budget constraints are milder than input constraints. The total costs are fixed but a firm may vary with different resources. Lump-sums or discretionary funds are types of expenditures from funding levels to operational levels.

We speak of *revenue constraints* when the firm's generated revenue must not exceed a set limit. This restriction is not frequently used, but in the Netherlands there is an example. Until recently, Dutch hospitals were bound to a budget. This budget could be earned by achievements (number of nursing days, number of surgical operations, number of laboratory tests etc.). Whenever a hospital exceeded the ex-ante budget, the excess was compensated for by adjusting the tariffs for procedures the following year. The Dutch hospital sector is now in a transitional phase, for which a hybrid system is applicable. For example, if a firm is subject to such a constraint, it will naturally try to limit its output in such a way that the revenue will be below the set limit. More output requires extra resources, but financially the firm will not gain anything. This type of constraint may incentivize inefficient behavior, something decision makers must be wary of and with proper research, answers to the extent of such inefficiency would be provided.

Most firms in the public sector have *profit constraints*. Limited profit is sometimes allowed in order to make future investments possible. The profits, or, the reserves when the profits are too substantial, may sometimes be skimmed. This means that firms may be inclined to work more inefficiently or increase the quality of the service. We can sometimes also speak of disguised profit maximization. (see § 3.3.5).

3.5 The optimal situation

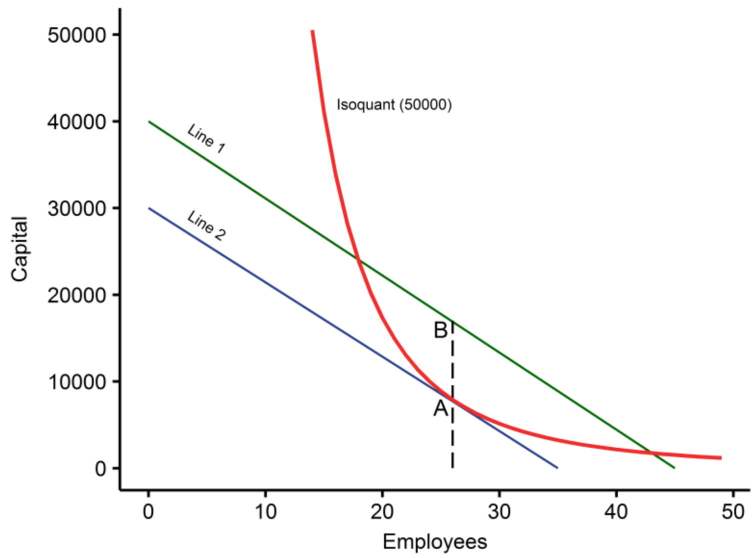
3.5.1 Technical optimum

Subsection 3.2.2 dealt with technical efficiency, focusing on the relation between the physical quantities of the resources and services. The assessment of production technology is dominant. The economic deliberations that are important in determining the input set and output set and their size were not taken into account. These economic considerations namely; the economic efficiency, which depends on the assumed economic behavior and the valid constraints are discussed here.

3.5.2 Economical optimum

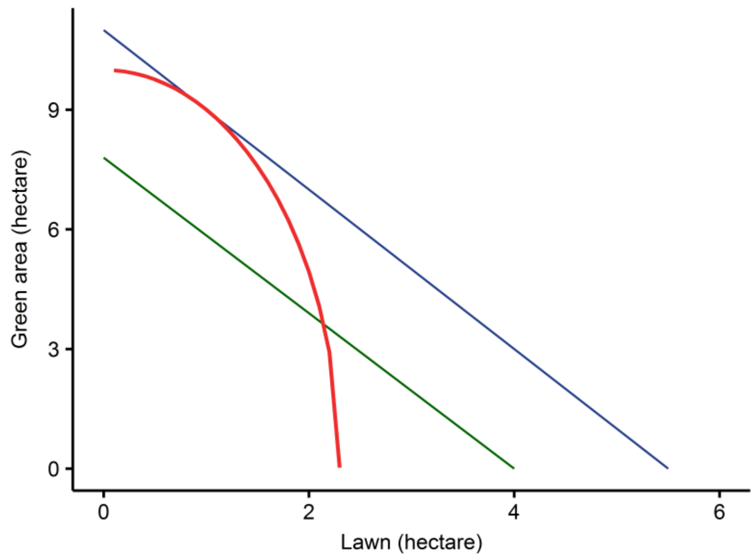
Figure 3-5 visualizes the optimal mix of capital goods and employees at a given output (=€ 50,000). A unit capital costs € 70, while an employee costs € 60,000. All combinations of capital goods and employees with equal costs are indicated by an "iso-cost" line (line 1 and 2). It should be noted that the "iso-cost" line is analogous to the budget constraint. If the budget is set at line 2, organizations cannot exceed this amount or geometrically produce above this line. Firms can produce at A, but not at B. The total costs at line 1 are € 2.8 million and at line 2 € 2.1 million. The firm now selects the combination that provides the desired output against the lowest costs. That is the point that meets the isoquant; in this case, a point on iso-cost line 2. Any other combination that provides the required output will automatically be located on a higher iso-cost line and therefore involves higher costs.

Figure 3-5 Optimum cost minimization and output constraints



A similar optimization can be deducted from revenue maximization at fixed resources. The red line in Figure 3-6 reflects all possible output sets (tilled public park and mowed lawn) at a given number of employees. This curve is known as the transformation curve. The blue and green lines are called iso-revenue lines; they reflect the output sets that generate the same revenue (in terms of money). The farther the iso-revenue line is from the origin, the higher the revenues are.

Figure 3-6 Optimum revenue maximization and budget constraints



Referring back to the park production example, a gardener receives € 2,000 per hectare of mowed lawn, and € 1,000 per square hectare of tilled public park. The optimum is reached at 0.9 hectare of lawn and 9.2 hectares of public park. The revenues then are €11,000.

The above is known as economic or allocative efficiency, thus referring to the choice of the mix of resources or services (or both). This means that this form of efficiency always relates to a money value, such as the costs or the revenues (or profits). When a firm has the opportunity to supply the same services against lower costs by adjusting the resources' mix (e.g. fewer staff members in favor of more material supplies), but does not do so, we speak of allocative inefficiency. The same holds for firms that can improve their revenues by shifting the services toward a more lucrative mix while holding resources constant.

3.5.3 Technical changes and economic optimum

Concepts such as technical, social, and institutional changes may affect at firms' productivity. This is not just about changes regarding quantities of resources, but also about changes regarding their combination. Some technical changes will lead to reducing the use of labor, and some will affect mostly large firms. There are three approaches to this concept:

- Hicksian neutral technical change;
- Input biased technical change;
- Output biased technical change.

The Hicksian neutral technical change refers to a development that equally influences the relation between services and resources. This means that the same services will be supplied with a proportional decrease in all resources (thus smaller labor forces and less capital are required in the same combination).

The input-biased technical change refers to a development that influences the mix of resources, for example in the direction of a much less energy-intensive production method as the price of energy increases vis-à-vis a suitable substitute.

The output-biased technical change indicates that the effect of change is dependent on the level of services. Some changes will be more effective in a large (or rather, small) firm or in a firm with a very specific mix of services, for example a highly specialized firm. These two forms of non-neutral technical change are known as *disembodied technical change.*

In empirical research, the difference between these various types of technical change is not always apparent. It is often assumed that there is a neutral technical change.

3.6 The influence of the environment

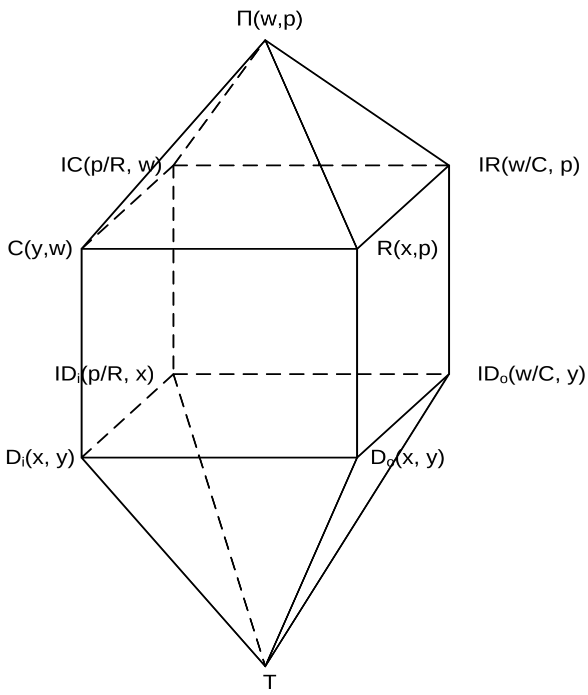
The environment's influence on the productivity of firms was introduced in Section 2.5. In modeling productivity, environmental factors must be explicitly addressed because the environmental factors influence technology. For example in a dry and warm climate, some cultivated plants do well due to intensive spraying. In a wet and cold climate, they may need more fertilizer, but much less water. The technologies for growing the same plant may vary just like the environmental factors. In terms of Figure 3.2, this means that the isoquant can have a different form, shift, or rotation. At given prices, this also means that another optimum is valid. Social environment implications include welfare considerations outside the strict cost-minimizing conditions, for example, schools in underprivileged areas may have different group sizes, have fewer individual contacts with students, and must employ more resources to ensure safety. Here, we also speak of substantially different technologies for schools in a different area based on population characteristics, also resulting in a different isoquant.

3.7 Specifications of economic relations

For each situation with certain economic behavioral assumptions and constraints, a distinct economic model can be deduced. Figure 3-7 shows a schematic overview of the most important economic models. They fit in a theoretical frame that, is known as the 'duality diamond'. This figure was originally presented by Färe & Primont (1995).

Every angle reflects an economic relation, with complimentary factors between brackets (w= prices of resources; p= prices of products/services; x= resources; y= products; I= indirect; D= distance; C= costs; R= revenue; Π = profit). Interested readers can also refer to the overview in the Annex of Chapter 3.

Figure 3-7 Duality diamond



Source: Färe and Primont (1995)

The lowest level in the duality diamond indicates the technology (T). T is a set of all the combinations of services and resources that are technically feasible. The level above is that of the distance functions (D), that is the combinations that are technically efficient. The back of the diamond shows the indirect distance functions (ID), namely the technically efficient combinations at a given cost budget or revenue goal. At one level higher, the front of the diamond represents the economically efficient resources and services mix (from the cost side or from the revenue side). The back stands for an economically efficient combination of resources and services at a given cost budget or revenue goal. (IC(p/R, w) and IR(w/C, p)). The top of the diamond features profits (Π). At given prices of resources and prices of services, all economically efficient resources and services mixes can be found here. Economically in this context relates to both the cost and revenue side.

When all the necessary data are available, all models can be applied and the inefficiencies calculated. These will most probably differ considerably, because each starts from its own perspective. This is because when assessing the productivity or efficiency, it is important to formally state the reference point.

3.8 The cost function

This chapter started with a production function, which is a formal representation of the frontier of the production technology. From Section 3.7 we know that there also are other representations which also include economic behavior and economic constraints. By far the most popular representation is the cost function (the left upper corner in Figure 3-7). The cost function reflects the mathematical relationship between cost on the one hand and services produced and resource prices on the other. This makes sense, since cost will rise as the produced output rise and/or the wages or other input prices increase. The cost function model has been applied empirically in thousands of studies. The main reason for this is that the required data for the model mostly are available. Further, the model has a strong intuitive appeal. The formal mathematical representation is $\boldsymbol{C} = \boldsymbol{c}(\boldsymbol{y}, \boldsymbol{w})$. Here we present the cost function in logarithms. The main reason for this is that it is common practice, mostly for econometric reasons, but also for reasons of interpretation. The parameters can be directly interpreted as elasticities, such as the scale elasticity.

$$\mathbf{ln}(\boldsymbol{C}) = \boldsymbol{c}(\mathbf{ln}(\boldsymbol{y}), \mathbf{ln}(\boldsymbol{w})) \qquad \text{(3-12)}$$

An example of a cost function with two outputs and two input prices, based on the Cobb-Douglas specification, is:

$$\mathbf{ln}(\boldsymbol{C}) = \mathbf{0.1} + \mathbf{0.6\,ln}(\boldsymbol{y_1}) + \mathbf{0.3\,ln}(\boldsymbol{y_2}) + \mathbf{0.3\,ln}(\boldsymbol{w_1}) + \mathbf{0.7ln}(\boldsymbol{w_2}) \qquad \text{(3-13)}$$

C = costs;
y_1 = produced services 1;
y_2 = produced services 2;
w_1 = price input 1;
$w2$ = price input 2.

Note that this expression is equivalent to:

$$\boldsymbol{C = e^{0.1} y_1^{0.6} y_2^{0.3} w_1^{0.3} w_2^{0.7}} \quad (3\text{-}14)$$

From the cost function we can easily derive a number of economic characteristics. Economies of scale, for example, follows from the cost flexibility:

$$\boldsymbol{cf = \sum_{m=1}^{M} \frac{\partial \ln(C)}{\partial \ln(y_m)}} \quad (3\text{-}15)$$

cf = cost flexibility.

This is the formal way. The mathematical expression says that every small relative change of the production of a service causes a small relative change in costs. The sum of all these small relative cost changes reflect the total relative change in cost. Let us take for example a change of 1% in each service. Assume that we start with y_1=2 and y_2=1 and for simplicity w_1=1 and w_2=1in the aforementioned example, then $C = 0.1 * 2^{0.6} * 1^{0.3} * 1^{0.3} * 1^{0.7} = 0.15157$. With an 1% increase in services produced y_1 = 2.02 and y_2 = 1.01, costs are equal to:

$C = 0.1 * 2.02^{0.6} * 1.01^{0.3} * 1^{0.3} * 1^{0.7} = 0.152935.$

The relative growth of C equals then $\frac{0.152935}{0.15157} - 1 = 0.0090 = 0.9\%$. This is lower than 1.0% implying that relative cost increase is lower than the relative rise in services produced. Therefore we have economies of scale.

If we use the mathematical expression (3-13) and apply this to expression (3-12) we will find the outcome 0.9 right away.

Another important characteristic is marginal cost. Recall, marginal cost reflects the extra cost when producing one extra unit of a certain service. The formal definition says that it reflects the ratio between extra cost due to an infinite small amount of extra production and the extra production. Mathematically:

$$\boldsymbol{mc^m} = \frac{\boldsymbol{\partial C}}{\boldsymbol{\partial y_m}} = \frac{\boldsymbol{C \cdot \partial log(C)}}{\boldsymbol{y_m \cdot \partial y_m}} \qquad (3\text{-}16)$$

mc^m = marginal cost of service m;
$\frac{\partial C}{\partial y_m}$ = partial derivative of cost with respect to service m;

In a less formal approach we use:

$$\boldsymbol{mc^m} = \frac{\boldsymbol{\Delta C}}{\boldsymbol{\Delta y_m}} \qquad (3\text{-}17)$$

With:
ΔC = change in costs;
Δy_m = change in service m delivered.

Equation (3-15) says that y_m should be increased by Δy_m. Then calculate the cost at $(y_m + \Delta y_m)$ and consecutively the increase in cost ΔC. The final step is applying (3-17).

Note that the difference between (3-16) and (3-17) lies in the amount of change. In (3-16) the change is very small, whereas in (3-17) any (extra) amount can be substituted. Obviously, (3-16) and (3-17) do not produce the same outcomes. Equation (3-16) is the formal correct one and can be easily calculated by taking the derivatives. For those without a calculus background, (3-17) can be applied. In case the chosen changes in the $y_m's$ are small (3-16) and (3-17) coincide as the next example shows.

If we apply (3-16) to the example, again in the point y_1=2 and y_2=1, then we find:

$$mc^m = \frac{C \cdot \partial log(C)}{y_m \cdot \partial \log(y_m)} = \frac{0.15157}{2} \cdot 0.6 = 0.045$$

The more intuitive approach is to take an extra small amount of y_1, let us say Δy_m= 0.1 extra. Then total cost will be:

$$C = 0.1 * 2.1^{0.6} * 1^{0.3} * 1^{0.3} * 1^{0.7} = 0.15607$$

The cost increase is then 0.0045 (=0.15607-0.15157). Since the production increase was 0.1, marginal cost equals $\frac{0.0045}{0.1} = 0.045$. As an exercise the reader should take Δy_m= 1 and does the math again.

3.9 Allocative inefficiency in education

A good example of identifying allocative inefficiencies in the public sector is given in Blank et al. (2012), who investigated whether there was excess management in secondary schools. Based on data on about 600 secondary schools, they examined whether there were allocative inefficiencies in schools and, if so whether these were the result of an excessive use of management and supporting staff. They use an advanced technique to reveal these allocative inefficiencies (for technical details see Blank, 2009).

The researchers measure the services of a school by means of the number of students and a number of quality indicators, such as pupil performance at junior high of the school and the twelfth grade, bearing in mind a correction for the original school career advice and the number of special needs pupils. The use of resources consists of, for example, the management, the teaching staff, the directly and indirectly supporting staff, and the material costs.

Table 3-3 Major aspects of the research by Blank et al. (2012)

Sector	Secondary education
Type of model	Shadow indirect output distance function
Data	Cfi dataset of school panel 2003-2006
Production	Number of students per type of education
Quality	Yield lower school corrected for school career advice and share of unprivileged students, yield secondary school per type of education.
Resources	Management teaching staff, direct and indirect support staff, materials
Environmental factors	Size of board, urbanization
Efficiency factors	Quality of staff
Economies of scale/scope	For some types of education, small schools have strong economies of scale; most schools have no economies of scale
Technological development	Productivity growth lies between 0 and 1.9% per year
Efficiency scores	Inefficiency varies between 0 and 45%

The research shows that there were no systematic allocative inefficiencies. Putting it differently, it means that the management of schools were not consciously aiming to employ excess employment of management or supporting staff. There were also scale effects. At small schools, there can still be economies of scale indicating that in the long run when expansion may be possible, these schools should be made larger. For most schools, however, the optimal size has been reached.

The technical inefficiencies vary considerably, namely between 0% and 45%. The autonomous productivity growth varies between 0% and 1.9% per year and per school type.

3.10 Requirements of economic relations

From the previous theory we can deduce economic relations. These economic relations can then be visualized in mathematical statements. These statements, however, cannot be randomly compared to one another. They should have properties that fit the underlying theory. A wage increase can, of course, never lead to lower costs at the same output level. In this section, we explain a number of these requirements. Most of the requirements are rather trivial, but they are not always easily converted into economic models. The requirements are often checked ex post, after the parameters of the economic model have been established via econometric methods. Checking the requirements is not always done (or done properly) in research. This may be a nice opportunity for readers of reports to check whether the researchers have done their job thoroughly.

First, we shall discuss the requirements of a cost function. These are intuitively most striking. The requirements of the other functions are similar and will be summarized at the end of the section.

A cost function needs to satisfy the following requirements (we provide the technical economic term between parentheses):

- Non-decreasing costs per product: when the production of a service increases (at an equal output level for other products), the costs cannot decrease (non-decreasing).
- Never decreasing costs: each random combination of services of two firms can never have lower costs than the firm that has the lowest costs (quasi-convex).

Examples of well-defined cost functions, that meet these requirements, are visualized in Figure 3-8.

Figure 3-8 Examples of correct cost functions

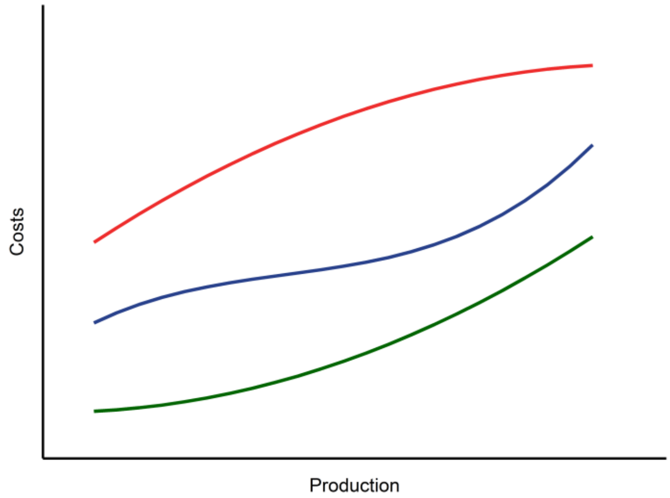
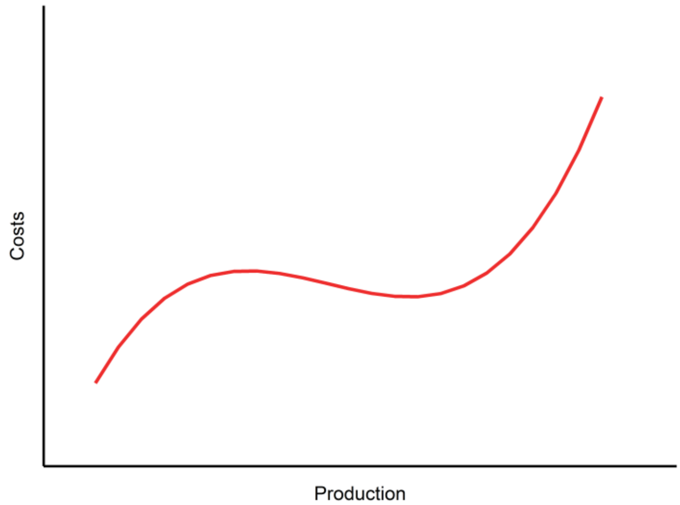


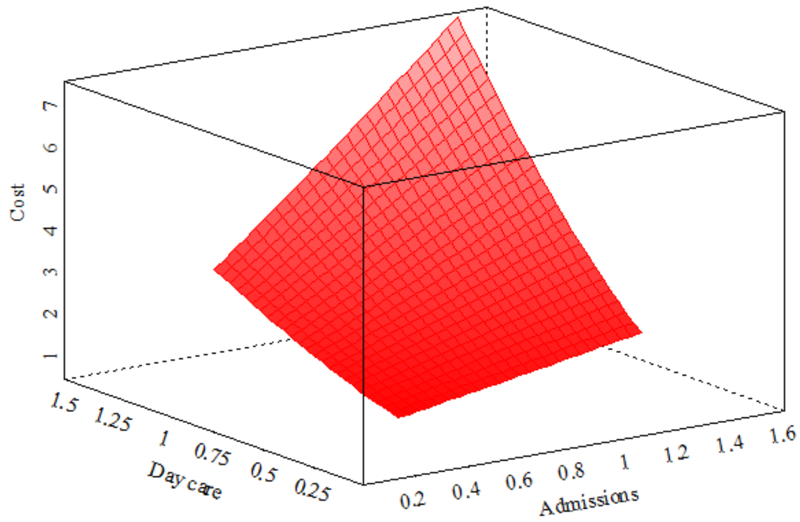
Figure 3-9 shows a false-defined cost function. In Figure 3-9 it is not difficult to find a combination of two points that have lower costs than the costs of both individual points. So it is not quasi-convex.

Figure 3-9 Example of a false cost function



When more than one product can be distinguished, it leads to so-called hyper planes that have a certain spherical shape. An example can be seen in Figure 3-10. The axes feature various different outputs and costs. It is clear that the lower plane has the mentioned qualities. The plane is non-declining in all directions and (as mathematicians put it so nicely) is, quasi-convex.

Figure 3-10 Relation between costs and output

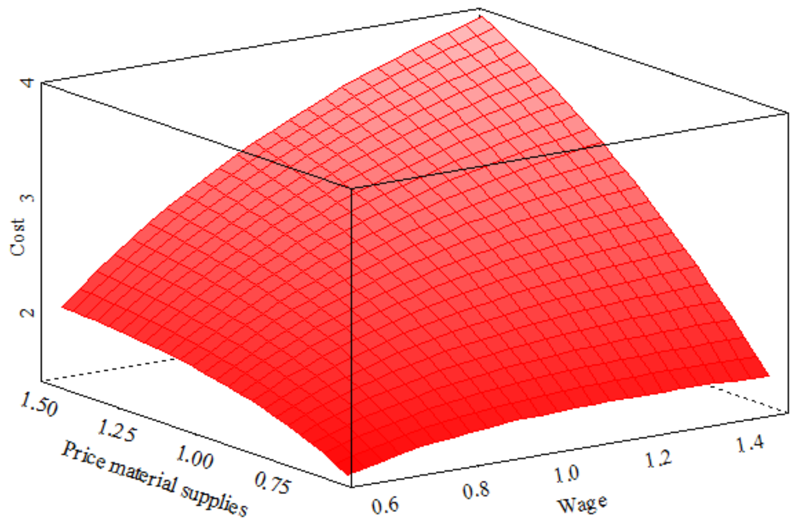


A cost function must also have a number of qualities with regard to the prices of resources (again the formal economic term is given parenthetically):

- Proportional cost increase: if the prices of all resources increase by a certain percentage, the costs will increase by a proportional percentage (homogenous of degree one).
- Non-decreasing costs at a price increase: when the price of a resource increases (at equal other prices) the costs cannot decrease (non-decreasing).
- Non-increasing cost growth at a price increase: the total costs cannot increase by a higher percentage than the original amount of a resource multiplied by the price increase of that resource (concavity). When half of the total costs consist of labor costs and the wages increase by 4%, then the total costs can maximally grow by 2% (50% x 4%). In most cases the cost growth will be less, because firms will try to compensate for the price increase by replacing the resource with another resource, for example substituting cheaper employees or advanced automation for costly labor. As indicated above, automation has been identified in the US for the loss of high paying manufacturing jobs. If increasing employment in well-paying jobs is a political objective, accounting for automation will need to be met with complementary labor skills, NOT substitute labor.

Figure 3-11 shows the cohesion between the prices of resources and the costs. The plane is non-declining and spherically shaped (also known as concave).

Figure 3-11 Relation between costs and prices of resources



For each function mentioned in the duality diamond a number of requirements need to be fulfilled; for those who have an interest, these requirements are summarized in the Annex of Chapter 3.

4 Collecting data

4.1 Data types

Research requires good data. Given the research question and the chosen approach, we could present a wish list of data variables:

- Services/outcomes/quality (if desired, with prices/values);
- Resources/instruments (if desired, with prices/values);
- Environmental factors.

Data can be collected at various aggregation levels:

- National level;
- Sectorial level;
- Business level;
- Department level;
- Personal level.

Readers should note that these levels for data collection correspond to the levels of analyses given in Section 2.

Research on national data mostly focuses on the influence of policy variables that in turn can be used in international comparisons. Countries can differ considerably regarding their regulation or policy for a particular sector and are therefore very suitable as analytical units.

Analyses of various economic sectors are sometimes also useful to measure the influence of policy variables. That various sectors are differently regulated or managed is also true. Aggregated data on sectors such as hospitals, education, and manufacturing will then be compared. At a business level, the data on individual firms are compared. They provide the most information about the production process effects, control, and strategy of firms.

In some cases, there is a need for insight into the functioning of parts of the business. This is mainly about parts that are present in many different firms, such as the administration, the human resource department (HRM), and the information and technology (IT) support. This creates the possibility to make comparisons between parts of the business that transcend sectors. We can think of comparing the IT services of banks, insurance firms, executive bureaucracies of social security, and the tax collector's office.

Another vital difference in collecting data concerns the dimension of time, including:

- Cross section;
- Time series;
- Panel.

A cross section reveals the productivity process at a moment in time such as a comparison of various firms during the same year. Other time units are also possible, for example months, quarters, or seasons. In general, cross sections are not suitable for measuring the effect of policy instruments or other exogenous factors, because they are the same for all firms at that time. In other cases, it happens that, experimentally, a policy rule is applicable to a number of firms, or that changes occur in a limited number of regions. In those cases, when the participating firms have been chosen randomly, an effect can be determined using the cross section approach.

Time series contain data of a certain unit over a series of years. In time series, data are always aggregated, such as the total output and costs of, for example, all hospitals combined. Time series are especially useful to illustrate effects on an entire sector, such as changes in legislation or technology. A common restriction of time series is a lack of variation in the data, which makes it difficult to determine the effects that one variable has on another. This may be because only a limited number of observations are available over the same time period. Another significant drawback is that it is often difficult to collect consistent time series. Stopping certain registrations or altered definitions of some variables in the past, make a fair comparison over time unreliable or invalid.

A panel is a collection of cross sections in time. It may concern the annual data of firms over a number of years. Panels combine the better of two worlds, for two reasons. The first is that the observations in the cross sections contain sufficient variations to calculate the effects of business features. The second reason is that because of the observations at various points in time, the influence of changes in exogenous factors, such as legislation and available technology, can be determined.

4.2 Efficient data collection

Research agencies often collect data via surveys. These surveys include exactly those questions that are relevant to the research. For many sectors, however, there are sufficient data available via regular registrations. In many sectors, there are registration obligations regarding finances, declarations, and other executive aspects.

Education in the Netherlands has a formalized type of data collection. For example, each student in the Netherlands has a so-called education number registered at the Information Management Group (IMG). At the IMG, each student file contains such personal data as name, date of birth, and address. Schools also contribute to this database by entering the type of education, year and any exam results of each student. The educational inspector then extracts data to determine who has failed a year or changed schools, in order to ascertain a school's performance. In order to be considered for funding, schools must be held accountable for their exploitation in annual accounts. The Education Implementation Agency (in Dutch: DUO) collects and checks these data. Since schools must pay their staff, extensive data on the type of job, seniority, and grades of teachers are registered by the school's administration. From this process, an elaborate set of data on students and their performances, as well as the number, mix, and quality of the staff, is available for each school.

4.3 Strategic data collections

When using an existing data collection method, it is important to check who collects the data, what the data are used for, and for what reason, who supplies the data, and what checks have been done to ascertain the reliability and the validity of the data.

The data collector should preferably work for an independent agency that has no special interest in the outcome. A special interest may exist when there is a direct link between the reported performances and financing. There will be an incentive for the firm to exaggerate its performances when registering them. In other words, agencies with a stake in data findings may cherry pick who they survey. Hence, independent collection and verification are crucial for maintaining data integrity.

The objective of the data collection is also essential. When data are collected for statistical purposes, the accuracy is not as important as when financing is linked to them. It is much better to use data from, for example, the tax collector's office or the IMG, as the data will been checked several times and provided with accountants' statements. Furthermore, there are sanctions for deliberately supplying false data.

4.4 Data sources

In most western countries, good data are available at the national statistics offices. National economics bureaus also provide data, as do sectorial bodies (e.g. on health-care). Interesting data at a country level are also available from the Organization for Economic Cooperation and Development (OECD), The World Health Organization (WHO) (http://www.who.int/topics/health_economics/en/), and Eurostat. Examples are the interesting data sets on health and education (OECD, 2012a, 2012b).

For example, the most important source of data in the Netherlands is Netherlands Statistics. It presents many data on its website www.statline.nl. This website primarily comprises time series data on a high aggregation level. For privacy regulation reasons, data on individual firms, households, and individuals cannot be found on the website. However, that does not mean that these data are not available for academic or policy purposes. It is possible to use, under very strict conditions, data on individual firms. There are many other associations that also collect data. Examples are Dutch Hospital Data (data on hospitals), the Union of Water Boards (data on water boards), and the Union of Water Firms (data on the distribution of drinking water). Examples in the United States are the Census Bureau, the Bureau of Labor Statistics, and health-care data sources such as Health Care Cost and Utilization Project (HCUP) at the Agency of Healthcare Quality and Research (AHQR). To find the appropriate data source – a good and simple place to start is google.com. Pubmed which has over 26 million citations https://www.ncbi.nlm.nih.gov/pubmed. BRFSS (Behavioral risk factor survey by Centers for Disease Control -- US) https://www.cdc.gov/brfss/ has data on individuals in each of the states in the US on health behaviors, self-reported health status, and chronic disease. Also from the CDC is a new data set that covers the health of 500 US Cities https://www.cdc.gov/500cities/. The American Hospital Association Data has information on all hospitals operating in the US:
http://www.aha.org/research/rc/stat-studies/data-and-directories.shtml/.

4.5 Wish list data

A wish list will then be the starting point in the quest for suitable data sources and the corresponding data. The researcher will be confronted with a data collection that deviates from the optimal collection. This means that the researcher must make concessions at one moment, and be inventive by combining data, often from various data sources, at another. The compromise sometimes leads to the realization that the aim of the research must be redefined or the outcome interpreted differently.

4.6 Processing and checking data

It is essential to process and check data. We can distinguish three aspects:
- Amount of data;
- Incorrect data;
- Missing data;
- Meaning of the data.

In general, only a limited number of variables are available for any analysis. In the case of education, data should include information about students repeating a year (per year), premature dropping out (per year), successful exam percentages, and the average exam grade per class. For a school in secondary education that offers the entire range of education from lower secondary vocational education to grammar school (senior high), this means that nearly a hundred quality indicators are available. For the analyses there is room for only a few quality indicators. Researchers must be inventive and preserve as much information as possible, while reducing the number of variables for the sake of parsimony. This is usually done by compressing the various indicators by calculating an arithmetic or geometric mean on either weighted or non-weighted variables.

Valdmanis et al. (2008) were able to include quality directly in the measure of US hospital care services by using an additional constraint in the linear programming problem to account for "bad" outcomes in care delivery. They found that 3% of additional inefficiency of the sample hospitals was attributed to quality deficiencies. This is an example of including quality indicators directly in the analysis of efficiency and productivity.

Table 4-1 Major aspects of the research by Valdmanis et al. (2008)

Sector	US Hospitals
Type of model	Output-based DEA with weak disposability of outputs
Data	American Hospital Association (AHA), Medicare cost reports, healthcare cost and utilization project (HCUP), Solucient (case-mix)
Production	Outpatient surgeries, births, adjusted admissions, other patient days
Quality	Failure to rescue, infection due to medical care, post-operative respiratory failure, post-operative sepsis (avoidance of the above)
Resources	Bassinets, RNs, LPNs, other personnel, other beds, acute care beds, residents/interns
Environmental factors	Not included
Efficiency factors	Teaching status, competition of markets (Herfindahl), system membership
Economies of scale/scope	economies of scale prevail (=1.234)
Technological development	Not included
Efficiency scores	Overall = 1.35; technical = 1.09; congestion = 1.03 (Note: 1 is efficient, >1 inefficient)

Practice shows that many data collections have quite a few irregularities, varying from typing errors and incorrect data entry, to missing figures. Special checks and standard procedures can be used to identify such errors.

Another problem is related to interpreting the data. It is important to distinguish between dynamic and static data. Dynamic data refer to a quantity in a certain time period, while static figures reflect the situation at a certain point in time. The best known example is the use of labor. The desired data for using labor is the number of working hours per year. Statistics and registrations, however, often mention the number of employees on 1 January or 31 December of a year. A static figure can considerably distort figures when there has been a considerable growth or shrinkage. Clearing static figures of consecutive years or seasonal effects will provide a better approach of the factual employment than a static figure. This difference is of vital importance especially in a productivity analysis, because output figures are virtually always dynamic (the number of supplied services in a year or the total revenue). Dynamic figures are linked to static figures in such a situation. Suppose that during a year the monthly supply of services increases and the employment of labor increases proportionally. On balance, nothing happens to the productivity. If the use of labor is measured on 1 January, then an unjustified productivity growth will be measured in that year.

Detection of outliers
Outliers are data points greater than three standard deviations from the mean. This means that outlier data are significantly different and if included would skew or bias findings rendering findings misleading. Some data can immediately be recognized as outliers, because they are a multiple of the value of observations at other firms (or any observation unit) or of observations at the same firms in different periods. Other outliers are immediately recognizable because they are 'physically' impossible. Think of a thirteen year old employee or a 100 m^2 total building surface. This type of error is often a consequence of registration error (the age of the employee turned out to be thirty-one) or of an incorrect interpretation of the unit of measurement (for example, in thousands instead of the actual unit).

The improbability of outliers shows not only in the value of the variable concerned, but also in relation to the value of other variables. At first glance, there seems nothing wrong with the provided staffing costs and the use of employees. A calculation of the staffing costs per employee, however, shows an amount that supersedes the collective labor agreement scales. Such an occurrence illustrates that there is something wrong either with the staffing cost data or with the employment of staff. In this case, the data are not plausible in relation to one another. These types of inconsistencies can go much further, because the data of three or more mutual variables are not consistent. The staffing costs per employee can be plausible, but not in relation to the average age or the functions of the staff members. Advanced techniques for detecting such outliers have been developed; for example, regression techniques can be used to detect, so-called influential observations. There are robust regression techniques that visualize outliers (Rousseeuw & Leroy, 1987).

Imputing missing data
Some types of missing data arise, because someone in the firm either has not filled out data on a questionnaire or has not collected data. Omitting this firm from the analyses is an option, but not a desirable one. The reason as to why this is not a good option is because the lack of this firm's productivity data is

not only a deficiency, but the contribution this firm could make to determine the best practice is also missing. In such cases, it is not uncommon to estimate the missing data via imputing. There are various imputation methods, such as the method of:

- The firm's average;
- The neighboring years' average;
- The reference average;
- Auxiliary regression.

The firm's average is derived by using the average value of a variable measured over all firms. If for example the data on illness percentage is missing, then the gap can be filled with the average absence percentage of the entire sector. Using this average limits the expected measurement error of this observation. This would be true assuming a normal distribution of absenteeism due to illness is present for all firms in the sample.

The neighboring years' average can also be related to the missing values in year *t*, whereas these are available for year (*t* -1) and (*t* +1). We can get a reasonably reliable estimate of the missing value by just averaging the value of two neighboring years.

The reference average is the average of the values of firms that are very similar to the firm with the missing value. But this may be limited if similar firms operation in different environments. For example, in the US some hospitals operate in states where Medicaid (insurance for the poor) was not expanded, these hospitals would have a lower efficiency rating not because of poor management, but because of state politics.

In an auxiliary regression, it is assumed that the variable (of which a value is missing) has a degree of consistency with variables of which the values are known. In statistical methods (linear regression), the consistency can be determined quantitatively. This consistency is then used to make an estimate of the missing value.

When applying imputation techniques, researchers must take into account that it is just an estimate and it may distort the results of the analyses. Applying the approach can be defended with the argument that the alternatives namely leaving out the variable or observation concerned will have a more profound negative effect on the reliability of the results. The following approach seems obvious: When the value of a variable is missing from a large number of observations, remove the variable from the analysis. A simple rule of thumb can be applied: *if the observation contains missing* *information on several variables, remove the observation. If this is not the case, use an imputation technique*.

When using this rule of thumb, the researcher must always be aware that a measurement error has been made, and that it will lead to unreliable estimates of certain effects or relations. In practice, this is occasionally solved by varying the imputed values and analyzing what the consequences may be. This results in an estimate of the order of magnitude of the extent of the inefficiency. This technique is known as *bootstrapping*, whereby the data are replicated to produce a multitude of samples so that errors in measurement can be ascertained.

Reducing data

Common sense plays an important role when reducing data. In fact, a researcher wants to use as few data as possible (parsimony) without losing the information value. Here are a few examples that illustrate this principle. To measure the use of resources, it is often considered adequate to distinguish a few types of employees and the use of material and capital (buildings, transport, machines). When differentiating types of employees, it is often sufficient to differentiate between management, professionals, and support staff. It is mostly not useful to make a more detailed differentiation, even though some organizations have as many as 200 functions. All too soon researchers find their data contaminated, because these functions have not been unambiguously defined and because there are considerable differences in function descriptions between firms.

A high degree of differentiation often only leads to the introduction of noise. Various functions can easily be aggregated by adding numbers of persons or full-time equivalents (FTEs). It is often also possible to add the different functions as weights, for example when there are various competence levels. For example, in hospitals registered nurses are integral to the actual treatment of a patient, whereas a licensed practical nurse conducts simpler patient care such as blood pressure, temperature taking, and patient bathing. The different labor costs levels of functions can then be used as weighting factors. In the US, hospitals can be compared by using diagnostic related groups (DRGs). Interested readers are referred to the following video on how to calculate DRGs: https://www.youtube.com/watch?v=z8rpSnqa7vE/.

Heterogeneous groups

Reducing data becomes more problematic when there is more heterogeneity among different data. A hospital supplies thousands of different types of treatment. Naturally, analyses will not include thousands of product indicators. Moreover, it will be difficult to add these product indicators. Open heart surgery is entirely different from treating a sprained ankle. Therefore weighting is often difficult to apply, because it is not known which costs can directly be linked to it or how many treatments of one sort can be equated to another sort. Assessing case mix indices as weights or a separate output has been used. In order to formally address the issue of heterogeneity, several diverse statistical approaches are possible:

- Differentiate between more or less homogeneous groups;
- Statistical clustering techniques;
- Information, inequality and concentration indices.

More or less homogeneous groups

Differentiating between more or less homogeneous groups is based on common sense. We can think of, in the case of hospitals, day admissions, admissions with a short hospitalization, admissions with a long hospitalization, and special medical treatments (such as the previously mentioned open heart surgery). The high degree of heterogeneity of all the services provided has now been reduced considerably, because it is about treatments with a similar care intensity. In the case of the police, we can differentiate among tracing violations of the law, minor offences and serious crimes. The remaining heterogeneity within the group does not necessarily lead to distortions, because the law of large numbers also applies. This law implies that the intensity of some product indicators may deviate from the average, but that negative and positive differences level each other out.

Statistical clustering

Factor analysis (http://www.statisticshowto.com/factor-analysis/) and principal component analysis (http://setosa.io/ev/principal-component-analysis/) are statistical clustering techniques. These clustering techniques are based on correlation analysis. Variables that strongly correlate bilaterally can be integrated into one group. Because of the strong correlation among the variables (within a group), it cannot be determined to what degree a variable does or does not contribute to the outcome in a productivity analysis. When such a case arises, it is convenient to include the so-called factor or principal component as the explaining variable. In that case, a factor consists of a weighted sum of underlying variables that strongly correlate. In many cases, the largest part of the total variation in dozens of variables can be bundled into three or four factors.

Information, inequality and concentration indices

These types of indices are based on the distribution of various services. Several measures, such as the Herfindahl index , the entropy index, and the Gini coefficient can be employed. Here, we will discuss the Herfindahl index, which is commonly used as a measure of market concentration in modeling markets from monopolies to higher competitive markets. The definitions of the entropy index and the Gini coefficient - both commonly measures inequality in wealth distribution, can be found in the annex of Chapter 4. The Herfindahl index equals the sum of the squares of service shares:

$$H = p_1^2 + p_2^2 + \cdots + p_m^2 \quad (4\text{-}1)$$

In which p_m is defined as:

$$p_M = \frac{y_M}{y} = \frac{y_M}{y_1 + y_2 + \cdots + y_M} \quad (4\text{-}2)$$

p_m = service share belonging to group m (m = 1, ... , M).
y_m = number of services in group m.

For the example given above this means that the Herfindahl index equals 0.63 (= $0.25^2 + 0.75^2$).

The absolute outcomes have no meaning concerning content. That is why it is an index. Only the mutual relation is meaningful. The outcomes of the indices usually lie between 0 and 1. The higher the index, the more concentrated the service in a smaller number of services. This concentration is an indication that the service deviates from the average. These firms often supply high quality or focus on “difficult” cases. No theoretical preference can be given in advance to any of the mentioned indices.

In the case of a limited number of groups, as in this example, it is obvious to include these groups separately in the productivity analysis. The technique is purely statistical and is only meant to somewhat responsibly aggregate differentiating groups.

5 Measuring productivity and efficiency

5.1 Introduction

In the academic research on efficiency of public services there is extensive literature, both technical and empirical (for example Blank, 2000; Blank & Valdmanis, 2008; Fried et al., 2008; Ozcan, 2008). The methods used can be divided into four categories: partial index numbers, stochastic frontier analysis, data envelopment analysis and semi-parametric analysis. Without going into technical detail we will briefly explain these methods. Readers wishing to pursue research in this area are referred to Färe et al. (1994), Coelli et al. (2005), Cooper et al. (2007) and Fried et al. (2008). An interesting overview of different techniques can also be found in Johnson & Kuosmanen (2015) and Parmeter & Kumbhakar (2014).

5.2 Partial index numbers

As indicated in Chapter 3, in order to measure relative inefficiencies among organizations, we need to find the efficient organizations or, in other words, the organizations defining the "best practice". A commonly used and relatively simple method is that of partial index numbers. This method indicates on which aspects the organization scores well. However, this method has some limitations. First, one needs to establish which aspects promote efficiency rather than reduce it. However, for some aspects this simply is not entirely clear and normative elements are added. Another problem with this method is its partial quality based on aspects or characteristics that have been determined by the organization. Nevertheless, it is not clear how we can deduce a general image of the organization's efficiency from all these scores. We will further illustrate this later on in this chapter. We also stress that some sort of super organization that scores the maximum result on all aspects cannot be used as a benchmark.

Example of partial index numbers
In Table 5-1, we show an example of partial index numbers. Index numbers have been allocated to four schools of higher secondary/pre-university education. At first glance, school A seems the least efficient. The costs per student are €7,000, whereas in other schools the costs are €6,000. School A

appears to be doing well as far as school results go. The efficiency for the upper grades of both higher secondary and pre-university education is higher than in most other schools. It appears from these indicators that small schools, by definition, have higher costs per student because of the scale effect, and that this school simply spends its resources very well. The *span of control* of this school is very small, which is why the management share is modest.

Table 5-1 Example of partial index numbers

Variables	*A*	*B*	*C*	*D*
Students	500	800	1.200	1.500
Costs per student (in €)	7.000	6.000	6.000	6.000
Management (in %)	3	7	8	10
Teaching personnel (in %)	60	65	55	70
Education support personnel (in %)	10	5	17	2
Support personnel (in %)	27	23	20	18
Share of higher secondary-students (in %)	50	60	40	20
Performance lower secondary (in %)	70	70	80	90
Performance higher secondary (in %)	90	80	90	70
Performance pre-university (in %)	90	80	75	95

Is school A doing very well or does it employ too many support staff members, or is that perhaps inevitable in a small school? School C has a very small educational staff and a relatively large number of support staff. The school results, however, do not suffer from this. School C, for example, scores better in the lower grades than schools A and B, whereas the higher secondary efficiency is equal or higher. Which then is the best practice? For comments on the use of partial index numbers, we refer interested readers to Blank & Lovell (2000).

5.3 Integral index numbers

In order to accommodate the objections against partial index numbers and to model the complexity of the output structure, various techniques to measure efficiency have been developed. These techniques have a common principle, and that is finding the *best practice*. See figures 3-1, 3-2, and 3-3 where the curves represent best practice.

Once the best practice is known, we can determine how far each organization is, from the best practice by using actual data. This reduces to one figure all relevant information about services, resources, and any environmental factors. This figure reflects the degree to which an organization can increase its productivity or decrease its resources. When all relevant information has been included, an integral score can be utilized.

In order to calculate this integral score, it is essential to identify the best practice. Determining the best practice is done using the actual data of organizations. The best practice can be deduced as a derivative of organizations that exist in real life, and not as a resultant of some sort of hypothetical super-organization that scores well on all possible aspects.

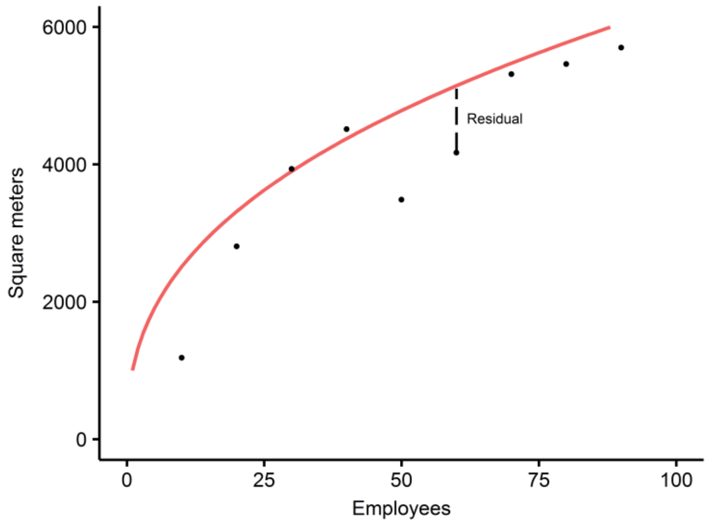
In general, there are three mainstream approaches to determining best practice: the stochastic frontier analysis, the data envelopment analysis and semi-parametric analysis. There are dozens of variants of these techniques, depending on the type of data and economic assumptions. Here, we introduce the three main types.

5.4 Stochastic frontier analysis

5.4.1 Principles of the methodology

In the stochastic frontier analysis (SFA) approach, the *frontier* is derived using a mathematical function (see, for example, equation (3-3)). The final form of such a function is determined by its parameters, which can be estimated using statistical methods. The purpose of the empirical research is to find adequate estimates for the parameters of the selected function based on the available data in order to determine efficient productivity. The selection of the mathematical function is dictated by theoretical and empirical considerations. The result is a smooth curve (as shown in Figure 5-1) which is derived from the observations on services and resources.

Figure 5-1 Stochastic frontier analysis



The deviation between the data point and the estimated curve is called the residual, and herein lays the solution to the problem. First, the residual contains measurement and specification errors. Measurements always suffer from a margin of error, and you will never find a mathematical statement that exactly reflects the technology. Second, the residual contains the degree of an organization's inefficiency. Measurement and specification errors can go both ways: They can both be positive and negative. The inefficiency, however, can only go one way. In relation to the output-based best practice frontier, this part of the residual is always negative. An inefficient organization will have a lower output than an efficient one holding inputs constant. When costs are the starting point, this part of the residual is always positive; inefficient organizations always have higher costs than the best practice. Econometric techniques can be used to separate these two parts of the residual (Aigner et al., 1977; Jondrow et al., 1982).

In the figures 5-2 and 5-3, we explain the measurement of efficiency in more detail. Figure 5-2 shows the residuals of all organizations in a histogram. The horizontal axis shows the values of the residuals in intervals (the size of 0.02). The vertical axis shows the number of times that a value appears in the corresponding interval. By carefully unraveling these residuals, we get Figures 5-3 and 5-4. Figure 5-3 shows the deviations as a consequence of the measurement and specification errors. This part of the residuals is both positive and negative, and has a so-called normal distribution with a mean around zero.

Figure 5-2 Frequency distribution of residuals

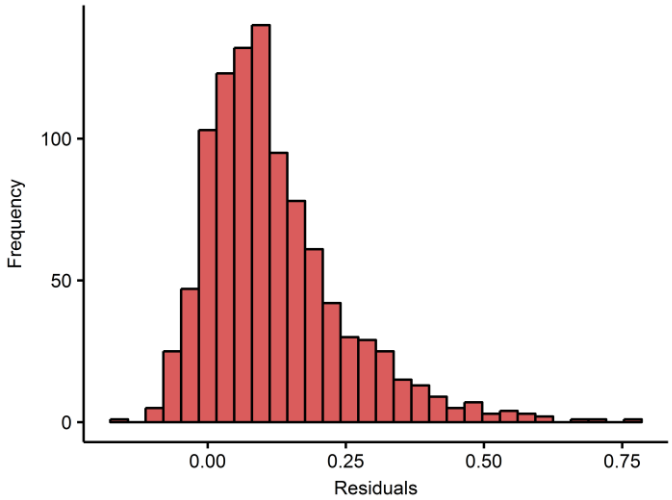


Figure 5-3 Frequency distribution of measurement and specification errors

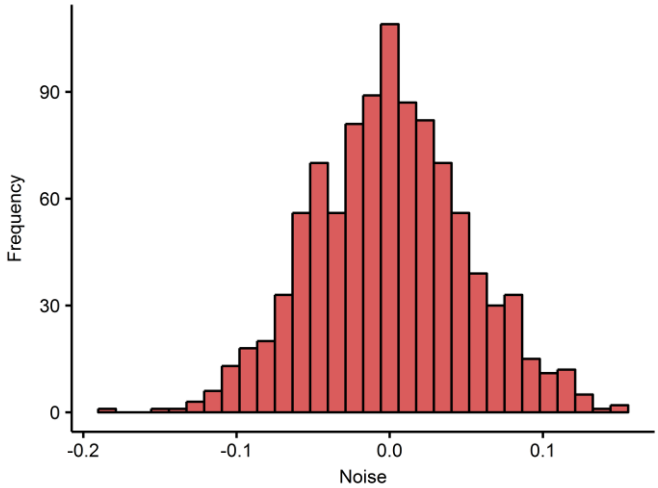


Figure 5-4 Frequency distribution of inefficiency

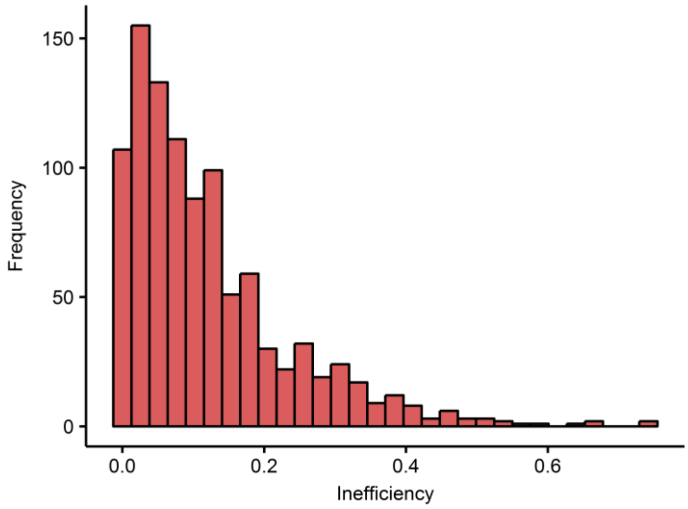


Figure 5-4 indicates the inefficiency. Here, all the residuals are greater than zero and also have another (assumed) statistical distribution. These can be considered: the exponential distribution or the half-normal distribution.

The curve in Figure 5-1 is now made up in such a way that the residuals should be as minimized as possible, whereas at the same time the curve should contain as many data points as possible. Note that not all points should be located below the curve; that is because of the measurement and specification errors. It sometimes appears as though a firm is doing much better than others, but this could be the result of, for example, an error in the data. This is an important fact, because the data envelopment analysis method (see § 5.5) envelopes all data points, including those with a considerable measurement error.

To determine the curve in such a way that the residuals are minimized, we need to apply econometric and statistical techniques. The simplest technique is when it is assumed that all residuals reflect inefficiencies. It start with conducting standard Ordinary Least squares (OLS) which will draw a regression line through the middle of all data points. The regression line will be shifted until all observations are below it (see Figure 5-2); This implies that only the estimated constant is adjusted in such a way that all the data points are enveloped. Note that all residuals have then become negative. This technique is called c*orrected ordinary least squares* (COLS). One problem with this approach is that it is sensitive to outliers. For example, if an organization reports a substantial output vis-à-vis other organizations, this organization will be a reference for others and, as a result, considerable inefficiencies will be measured. Note that even with the imposition of explanatory variables, this outlier may still exist. Therefore, reviewing and double checking the values in the data are warranted in order to determine measurement error.

Another approach is the *modified ordinary least squares* (MOLS). This technique assumes that the residuals are distributed in a manner that fits a situation that only allows positive residuals (see Figure 5-5), which may or may not mitigate the COLS outlier problem. These techniques are somewhat dated and, given all the available software and increasingly fast computers, there is a strong preference for using the more complex SFA.

A benefit of using frontier approaches comes with the increased availability of panel data. Because governmental and non-profit organizations are not strictly cost minimizers or profit maximizers, making these assumptions leads to misspecification, because considerations such as social goods cannot be explicitly included in such a strict economic framework. Hence, these panel data techniques assume that inefficiencies are organization-specific (and therefore constant in time) or follow a certain pattern in time. The idea behind this is that a poorly organized firm will not do much better the following year particularly without either market or budgetary constraints. It is clear that this assumption is difficult to maintain over a longer period of time, because poor performances will, at some point, be identified and it is hoped a reorganization will improve efficiency. In the long term, cyclic patterns of efficiency will be sometimes assumed. This/these constant factor(s) per firm is/are easy to determine and is/are an indication of the inefficiency of the firm concerned. For this purpose different variants have been developed. These techniques are common under the name *fixed effects* and *random effects* models. In a fixed effects model, the observed independent variables are treated as though they were non-random. In the random effects model, independent variables are treated as though they result from random causes. For example, hospital ownership can be considered in a fixed effects model, because it does not change and arises from a choice made by relevant decision makers. For an excellent overview of all possible techniques, we refer to Chapter 2 of Fried et al.(2008).

5.4.1 Limitations of SFA

One of the major limitations of SFA is that the researcher must, a priori, choose the mathematical relation between costs and independent variables. Even though several mathematical relations exist, it still means that the choice determines its form.

Another choice to be made concerns the form of the statistical distribution of the inefficiencies (see Figure 5-4). Again, there are several possibilities. Different research also shows that this choice may influence the outcome without making clear which option should be preferred.

In practice, it appears that estimates using SFA are not always successful. There are numerous examples where the residuals are entirely attributed to distortion or inefficiency. We then encounter so-called convergence problems. The computer software is not able to find some good solution and picks one of the extremes as the outcome.

Applying SFA to multiple equation models requires the use of sophisticated econometric modeling techniques such as Bayesian statistics and Monte Carlo simulations (see for example Kumbhakar & Tsionas, 2005).

5.5 Incentives and the regulation of airports

In this section, we provide an example of a SFA application. Martın et al. (2009) conducted empirical research on the influence of regulation on the efficiency of airports operating in Spain. Based on data on the 37 airports in Spain over a period of seven years, they estimated a model with a cost function and cost share functions.

The efficiency scores of the various airports vary considerably (from 15% to 26%). Especially small airports have scale inefficiency. Based on estimates of marginal costs, the researchers also state that the finance structure is out of line with the cost structure and that there is cross-subsidization. They also claim that the lack of sufficient stimuli is an important cause of the observed inefficiencies.

Table 5-2 Major aspects of the research by Martin et al. (2009)

Sector	Airports
Type of model	SFA
Data	Panel data of 37 airports in the period 1991-1997
Production	Number of air traffic movements, weight of cargo (0.1 * passengers + tons of cargo)
Quality	Not mentioned
Resources	Labor, materials and capital
Environmental factors	None
Efficiency factors	Not mentioned
Economies of scale/scope	Average airport faces economies of scale
Technological development	Average of 1.1% per year
Efficiency scores	15-26%

5.5.1 Single equation models and multiple equation models

Most studies work with a single equation model. A production or cost function is specified and its parameters are estimated. In Chapter 3, we showed that by applying certain assumptions, such as cost minimization, the optimal resources mix can be deduced. This optimal mix of resources is also dependent upon resource prices or service levels. This dependence can also be represented by an equation. The equations for costs and optimal use of resources have the same parameters (because of their theoretical dependence). The model's parameters can be estimated more accurately, because the researcher in fact uses more information. However, estimating the systems is far more complicated, and this deters many researchers from choosing this approach. Complicated modeling may also be inaccessible to policy makers who lack a statistical background.

5.6 Mergers of drinking water firms

An example of systems' analysis is given in Dijkgraaf & Varkevisser (2007) who researched the effects of mergers on the costs of water distribution firms that are responsible for transporting clean water to homes. They used the data of Dutch water distribution firms to analyze whether merged firms have lower costs than non-merged firms. Lower costs are to be expected when merged firms have economies of scale. They also researched whether lower costs occur because merged firms are better capable of eliminating technical or allocative inefficiencies.

Dijkgraaf & Varkevisser (2007) show that merged firms are no more efficient than non-merged firms. They could not find any convincing evidence for economies of scale either. However, economies of scale cannot be excluded based on these data. It appears that there are diseconomies of scale for smaller firms. That means that an increase in scale will lead to a decrease in costs.

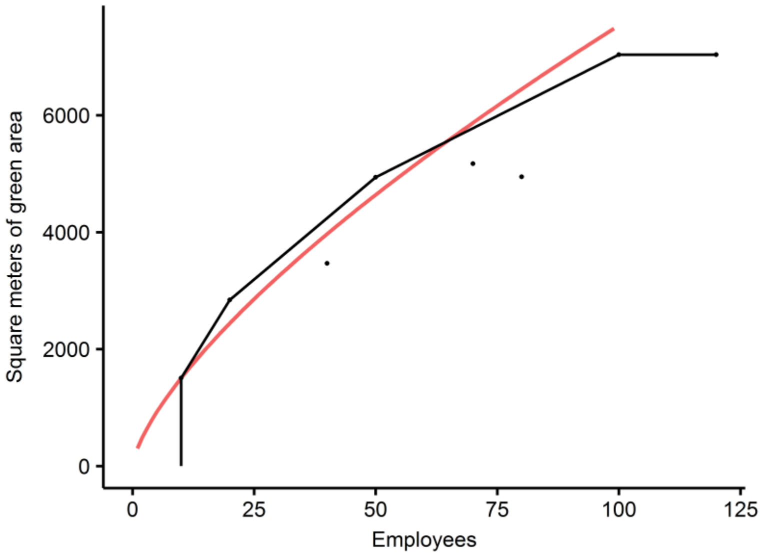
Table 5-3 Major aspects of the research by Dijkgraaf and Varkevisser (2007)

Sector	Water distribution
Type of model	Cost function
Data	Panel 1992-2006
Production	Sales, connections, network length
Quality	Not mentioned
Resources	Staff, capital
Environmental factors	Soil stability, purification efforts, age of installation
Efficiency factors	Not mentioned
Economies of scale/scope	Diseconomies of scale for small firms
Technological development	Not mentioned
Efficiency scores	Not mentioned

5.7 Data envelopment analysis

Data envelopment analysis (DEA) is a technique based on linear programming. This technique is derived from early production work by Farrell (1957) and Debreu (1951) and was later formalized using linear programming techniques (see e.g. Banker et al., 1984; Charnes et al., 1978a; Färe et al., 1986).

Figure 5-5 Data envelopment analysis



The objective of this approach is to envelop the data points as closely as possible, and to produce the best practice frontier by linking together several line segments. When DEA is applied to Figure 5-1 we get Figure 5-5. The (dark) blue envelopment is now the best practice and because of its linear combinations depict the frontier as connected segments which deviates from the best practice as defined by the SFA (green line, the smooth curve).

Overall efficiency can be decomposed into its component parts. Allocative efficiency (AE) which measures the right mix of inputs (or outputs). Technical efficiency (TE) which measures the right amount of inputs (or outputs). Scale efficiency (SE) which measures scale efficiency and finally congestion (CE) which relaxes the strong disposability of inputs (outputs) and provides information on economic bads that detract from overall efficiency.

By limiting the details of DEA to the most elementary parts there are two ways to clarify how DEA works. The first method starts from the output concept (primal approach); the second method starts from the output structure or the technology (dual approach).

5.7.1 Primal approach

In the primal approach, efficiency is linked to the concept of productivity. Productivity is defined as the relation between services and resources. In the simplest case of one product and one resource the productivity equals (see also § 2.2):

$$\boldsymbol{Productivity} = \frac{\boldsymbol{y}}{\boldsymbol{x}} \qquad (5\text{-}1)$$

y = service level;
x = resources.

Whenever there are multiple products and resources, we cannot use this simplified mathematical ratio unless we add more information or apply a more sophisticated approach. One option is to weight the different products and resources. The productivity then equals the ratio of the volume of the output to the volume of the input. A special case is the situation in which the prices of products and resources are available and can be used as weighting factors. The productivity then indicates the revenue per cost unit, which is also known as profit margin. So, the general expression for productivity is:

$$Prod_o = \frac{p_1 y_{1o} + p_2 y_{2o} + \cdots + p_M y_{Mo}}{w_1 x_{1o} + w_2 x_{2o} + \cdots + w_N x_{No}} \quad (5\text{-}2)$$

$Prod_o$ = productivity of organization o;
p_m = weight of product m;
w_n = weight of resource n;
o = individual firm being assessed.

There are two problems related to predetermining weights: The chosen weights may be arbitrary, and the weights for the different firms may be fixed. An alternative is to apply a different set of weights for each firm.

A much better option is to use a data-driven approach to empirically determine the weights. From this weighting approach, we begin to speak of efficiency. The concept of efficiency has a number of strict conditions and is therefore a more limited concept than productivity.

As indicated, the DEA method assumes that each firm values its output and input in a certain way, and that the firm's management tries to achieve optimal output based on this. We must therefore establish a set of weights in such a manner that the efficiency of each firm being measured is maximal, setting the figure of efficiency between 0 and 1. The set of weights for this firm should therefore be determined in such a way that all other organizations also result in a value between 0 and 1. In fact, the organizations all get the benefit of the doubt. Any other choice of weights will lead to a lower score. By optimally determining the weights per firm, managers can never "complain" that, for their firm, the weighting was wrong or that they are being assessed against an impossible standard. Furthermore, we assume that all different products positively contribute to the overall production; that is that all weights must be positive.

A solution to the above would lead to an infinite number of possible sets of weights. When all weights, for example, equal 1, the solution for which all weights would equal 2 or 10 would also be valid. The productivity would show the same outcome in all cases. Everything in both the numerator and denominator would be increased by the same factor. In order to avoid an infinite number of solutions, the weights are selected in such a fashion that for the observation $w_1x_1 + w_2x_2 + \ldots\ldots + w_Nx_N = 1$ is valid as well. In other words, we assume that the volume of the resources equals 1. This is a sufficient condition to generate a unique solution. At the same time, it is also possible to restrict the production volume. This is purely a technical condition. The absence of this restriction in the equation would confuse the optimization algorithm.

The problem for the researched firm *o* can then be defined as follows:

maximize *prod*

under the conditions:

(1) Each firm receives a score of less than or equal to 1. There are *I* firms. This condition must then be included for each of the *I* firms (numbered with *i* = 1, 2,..,*I*).

$$\frac{p_1 y_{1i} + p_2 y_{2i} + \cdots + p_M y_{Mi}}{w_1 x_{1i} + w_2 x_{2i} + \cdots + w_N x_{Ni}} \leq 1 \quad (i = 1, \ldots, I) \qquad (5\text{-}3)$$

(2) To create a unique solution the total input value of the researched firm is set at 1.

$$w_1 x_{1o} + w_2 x_{2o} + \cdots + w_N x_{No} = 1 \qquad (5\text{-}4)$$

(3) All weights must be greater than or equal to 0.

$p_1, p_2, \ldots, p_M \geq 0$
$w_1, w_2, \ldots, w_N \geq 0$

The problem mentioned above has the structure of a so-called linear programming model and can easily be solved with linear programming techniques. The model must be solved separately for each firm.

Box 5-1 Linear programming

Linear programming is a mathematical discipline that focuses on determining the outcome of the optimization of a certain goal, such as maximizing profits, under certain constraints. To a farmer the constraints may be the available land or the maximum amount of organic fertilizer. The concept of programming is related not to developing computer programs, but to coming up with efficient algorithms in order to solve the problem. The most common algorithm is the simplex method, developed by the American mathematician George Dantzig in 1947.

To illustrate the linear programming approach we use the following example. In Table 5-4, data on inputs and outputs used by the municipalities describe the production of the park-keeping service mentioned in Chapter 3. The outputs are defined tilled parks (in hectares) and mowed lawns (in hectares); the resources (inputs) are the number of employees. The last five columns indicate the outcome of Equation (5-3), which uses the weights as indicated above the columns. When the set of weights equals (1,5), the park surface is multiplied by 1 and the lawn surface by 5. For municipality 10 this means that the total output volume equals 1x2+5x5=27. Municipality 10 then has a productivity of 27/3=9.

Table 5-4 Example of the solution to a primal problem

Municipality	*Public garden*	*Lawn*	*Employees*	*Weights*				
				1,1	*1,2*	*1,5*	*2,1*	*5,1*
1	1	1	1	2.00	3.00	6.00	3.00	6.00
2	2	0	1	2.00	2.00	2.00	4.00	10.00
3	0	2	1	2.00	4.00	10.00	2.00	2.00
4	1	2	2	1.50	2.50	5.50	2.00	3.50
5	2	1	2	1.50	2.00	3.50	2.50	5.50
6	2	0	2	1.00	1.00	1.00	2.00	5.00
7	1	3	2	2.00	3.50	8.00	2.50	4.00
8	3	2	3	1.67	2.33	4.33	2.67	5.67
9	2	2	3	1.33	2.00	4.00	2.00	4.00
10	2	5	3	2.33	4.00	9.00	3.00	5.00
maximum				**2.33**	**4**	**10**	**4**	**10**

Table 5-5 Efficiency scores, example Table 5-4 (in percentages)

Municipality	*Weights*					*Efficiency*
	1,1	*1,2*	*1,5*	*2,1*	*5,1*	*scores*
1	86	75	60	75	60	86
2	86	50	20	100	100	100
3	86	100	100	50	20	100
4	64	63	55	50	35	64
5	64	50	35	63	55	64
6	43	25	10	50	50	50
7	86	88	80	63	40	88
8	71	58	43	67	57	71
9	57	50	40	50	40	57
10	100	100	90	75	50	100

To place the outcome in a relative perspective, we adjust Table 5-4 by dividing each outcome by the outcome of the best practice park services, which is why the last row also includes the outcome of the best performing services. Each number will then be divided by the corresponding maximum of the column. Each column now contains at least one municipality that has a score of 1 (or a 100% score)-(Table 5-5). The outcome here is shown in percentages.

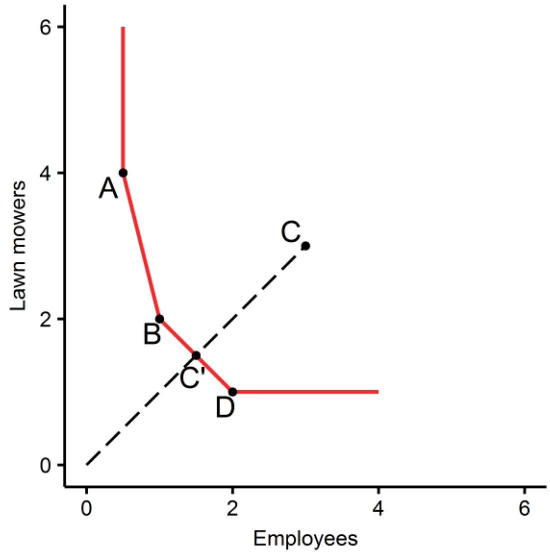
Each municipality can now select its best score. In Table 5-5, we show that municipality 1 realizes its best score (=86%) by opting for weights 1 and 1 (principle of the benefit of the doubt). Municipality 8 also realizes its best score at weights 1 and 1, but consequently does not have an efficiency score higher than 71%. These best options are then summarized as efficiency scores in the last column. Three municipalities are efficient (2, 3, and 10), the others have an efficiency of 86%, 64%, 64%, 50%, 88%, 71%, and 57%.

Suppose an a priori weights system had been chosen, for example 1 and 2, then it would have looked different. There would have been two efficient municipalities (municipalities 3 and 10), and municipality 2 would have had an efficiency score of 50% (instead of 100%). Policy-wise the implications may be profound because the wrong combination of outputs and inputs may be utilized, leading to higher costs.

5.7.2 Dual approach

The above-mentioned problem can be redefined in the dual approach. In the dual approach, each firm is compared with the best practice of other firms or with certain fictitious firms that are a combination of existing firms. Are there firms or combinations that have a better performance? In other words, are there firms or combinations thereof that produce equal amounts of services with fewer resources? The degree to which the observed firm can decrease its resources in order to reach the level of the better performing firm, is called the firm's inefficiency. Firms whose performance is better than other firms are considered to be producing efficiently, defining the best practice frontier.
The problem takes the organization that is observed as the starting point and tries to maximally reduce the resources without affecting the output (see Figure 5-6.)

Figure 5-6 Firms and their best practice



The quantities of resources (e.g. employees and lawn mowers) are indicated on both axes of Figure 5-6. Points A, B, C, and D are municipalities that all realize the same output. They each use different quantities of employees and lawn mowers. Let us take firm C as the starting point. Firm C can reduce both resources proportionally. There is an imaginary firm on line BD (C') that generates the same output as C but with less input. The relation between these virtual resources employed by C' and the actual resources of C (= OC'/OC) is known as the technical efficiency of firm C. A reduction of resources is not possible for A, B, and D, because there are no firms or combinations of firms that generate a certain output with fewer resources. In that case, we say that A, B, and D are at the best practice or the frontier. Please note that the input set and isoquant, contrary to Figure 3-2, cannot be presented as a smooth curve, but as a collection of line segments linked together.

The DEA, as modeled in the example above can easily be summarized in mathematical statements referred to as simple optimization problems. For this statement, we take two resources (x_1 and x_2) and one product (y). For the sake of convenience, there are three data points (B, C, and D) for which we will calculate C's technical efficiency. We will now try to find the smallest possible fraction θ by which we can multiply the resources x_1^C and x_2^C. This results in the following expression:

Minimize θ

At the same time we must see ensure that we can continue to achieve the same output (y^C). That means that there must be a firm or a combination of existing firms that produces at least as much with fewer resources. Or:

$$\lambda^B x_1^B + \lambda^D x_1^D \leq \theta x_1^C$$

(the firm or combinations of firms uses as much or less than θx_1^C)

$$\lambda^B x_2^B + \lambda^D x_2^D \leq \theta x_2^C$$

(the firm or combinations of firms uses as much or less than θx_2^C)

$$\lambda^B y^B + \lambda^D y^D \geq y^C$$

(the firm or combinations of firms produces as much or more than y^C)

The firm or combinations of firms is therefore determined by also defining the λ^B's and λ^D's. This is also a linear programming problem that can be solved in Excel.

The dual approach also provides the opportunity to formulate numerous other problems. Whereas it is not within this book's scope to delve into detail, it is useful to mention a few of these alternatives.

Other issues that can be examined via DEA include scale, disposability, and changes over time. The original Charnes, Cooper and Rhodes-model (CCR) assumes constant returns to scale. That means that an increase in all resources by a certain percentage will yield a proportional growth of the output. It is also possible to formulate the above-mentioned problem in such a way that we get variable returns to scale. This means that the output in such a situation grows more or less than proportionally. This model is known as the VRS model (*variable returns to scale*), which was first introduced by Banker et al. (1984). Comparing the outcome of the CCR model and the VRS model provides an insight into scale inefficiencies.

The previously mentioned model was defined in terms of reduction of input at a given output, called the input orientation. The model can also be defined in terms of increase in output at a certain input (output orientation). Whenever price data on resources or on the supplied products or services are available, the model can be redefined in terms of cost or revenue (cost or revenue model).

Comparing the outcome of these models with those of previously mentioned models will indicate inefficiencies as a consequence of an inaccurate input set or an inaccurate output set. This can be used to detect a relatively large overhead within a firm. This form of inefficiency is known as allocative inefficiency.

In some cases, outputs or inputs are not strongly disposable, which means that the reduction of an economic output or input is not costless. For example, pollution is an economic bad that increases as production increases. A researcher or analyst using the DEA approach has the capability to relax the strong disposability of outputs on the amount of pollution produced to gauge the social welfare loss of pollution.

It is possible that, in the course of time, the determined *best practice* will shift as a result of technical or social changes. This shift over the years can also be mapped by applying DEA for the different years and comparing the outcomes in a special way. For this purpose, the Malmquist indices are calculated (Caves et al., 1982; Coelli et al., 2005; Färe et al., 1994). This index measures both the shift of the *frontier* and the change in distance of firms compared with the corresponding *frontier* (change in the technical efficiency).

A strict variant of DEA is referred to as the free disposable hull (FDH) method which makes exclusive comparisons based on the actual data of firms (and not on linear combinations of firms). This means that an organization is inefficient when another organization produces more (or the same) with fewer resources (input orientation), or when a firm produces more at an equal (or minimal) use of resources (output orientation). The FDH method will show fewer inefficiency differences than the DEA method. In this method a unique firm that has no reference firms will be efficient by definition. A mathematical formulation can be used with λ restricted to the values 1 or 0 and summing up to 1. This means that the linear programming optimization is searching for the smallest θ under the constraint that there still exists one firm that is producing as least as much as the firm under investigation. Further, the best practice firm uses an amount of inputs that does not exceed θ times the inputs used by the firm under investigation.

Dervaux et al. (2009) used a variant of the FDH model referred to as the robust FDH to analyze the use of resources by patients in intensive care units in Paris. FDH was used in this case because, traditional DEA would have been inappropriate, as patients cannot be radially contracted (in other words, a patient cannot be compared to half of one patient and half of another patient) as is the case with decision-making units.

Table 5-6 Major aspects of the research by Dervaux et al. (2009)

Sector	Intensive care units (ICUs) in Paris
Type of model	Robust free disposal hull
Data	Patient-level data at 25 hospitals and ICUs the Paris region 2000
Production	Probability of death – discharge status
Quality	Not included
Resources	Treatment regimens and length of stay
Environmental factors	Diagnoses categories, SAPSII score, model of entry, surgical vs. medical
Efficiency factors	Not included
Economies of scale/scope	Not included
Technological development	Not included
Efficiency scores	0.70

5.7.3 Advantages and disadvantages of DEA

Each method has its limitations. A disadvantage of using DEA is that it implicitly assumes that all variations in the performances of different firms are a consequence of inefficiencies. Any irregularities (measurement errors) in the empirical data, as well as differences in the quality of output are interpreted as inefficiencies. The inefficiencies can be dramatically different not only for the firms that have these outliers, but also for the other firms. Another disadvantage of DEA is that in most variants the method only measures firms that have comparable output levels. The method must therefore be applied to by using comparable organizations. Very large firms will easily be labeled efficient, because there is no other firm with a similar output level. If constant returns to scale is imposed, a large firm can be compared to a scaled up version of a smaller firms using the same technology (same proportion of inputs and outputs). Also, when more products are distinguished, the number of firms with a comparable output level will decrease, and therefore the number of efficient firms will increase. And finally, the method often leads to an exaggeration of the scale effect, because smaller firms are often labeled as scale efficient. If one small firm has a very efficient management, it will become a reference for all firms. The previously mentioned sensitivity to outliers is especially valid here.

On balance, DEA is a method that does justice to the *multiple input, multiple output* character of organizations. DEA is a comprehensible method for which a number of standard programs are available. That is why the calculations can be executed and interpreted relatively easily and with better data and appropriate caveats, decision making can be guided by the findings.

5.8 Forestry management

Here, we present an example of DEA assessing scale, technical and allocative efficiency. Oude Lansink et al. (2008) present the results of an efficiency research of forestry in the Netherlands. They present the effects of a number of qualities of forestry that influence the efficiency. The outcome has been summarized in Table 5-7.

The research shows that efficiency differences between forestry firms are considerable. There are substantial differences in the size of the firms and the input sets. Both aspects largely determine the differences. The scale inefficiency is 23%, the technical inefficiency 28%, and the allocative inefficiency 37%.

A striking aspect of this research is that, even though the researchers had access to the data over a four year period, they did not use it as such. Instead, they average all data over four years and therefore analyzed the data of a cross section. The background of this approach is that production in nature is a long-term process. Efforts in one year only lead to new output in the future; therefore analyses based on years would provide a distorted image.

Table 5-7 Major aspects of the research by Oude Lansink et al. (2008)

Sector	Forestry
Type of model	Data envelopment analysis
Data	LEI company-information network 2001-2004 (84 companies)
Production	Revenue wood sales, subsidies, remaining revenues (recreation, yacht lease, Christmas trees)
Quality	None
Resources	Total costs of labor, costs of outsourced work, surface of forest
Environmental factors	Not included
Efficiency factors	Cut down trees one selves, size class, region, urbanization
Economies of scale/scope	Scale inefficiency = 23%
Technological development	Not included
Efficiency scores	Technical = 28%; Allocative = 37%

5.9 Non-parametric frontier analysis

Stochastic frontier analysis and Data Envelopment are by far the most popular methods to estimate productivity and efficiency. As argued earlier, both methods have their pros and cons. To summarize, the critics of SFA focus on the required mathematical specification of the model and the distributional assumptions about the efficiency component, whereas DEA does not require these. The critics of DEA focus on the absence of a stochastic component and the difficulty of controlling for environmental variables and deriving economic features like economies of scale and scope and input (or output) substitution, whereas SFA can easily include these.

Only in recent years has there been a tendency in the literature to try to combine the best of both worlds. Kuosmanen (2008) developed a technique that converts a DEA formulation into a stochastic formulation that can be estimated by maximum likelihood techniques. This technique is called STOchastic Nonparametric Envelopement of Data (STONED). We will not elaborate on this any further.

Another appealing approach was developed by Fan et al. (1996), who used standard kernel methods based on maximum likelihood. A kernel method is nothing more or less than calculating a local (weighted) average. In practice, we will not be able to identify identical firms, but instead a number of similar firms. To take this into account very strong similarities, they get a large weight, whereas firms not as similar get a small weight. The weights are being used to derive the weighted average of cost. Note that this technique produces for each observation a weighted average, depending on a different set of similar firms (and different set of weights).

There also exist an extension of this method called local linear least squares. In this case a (weighted) regression method is being locally applied to a very simple linear equation (note we left out the input prices for the sake of simplicity). For instance, in case of two services the cost equations could look like this:

$$\boldsymbol{C = a + b_1 y_1 + b_2 * y_2} \quad (5\text{-}5)$$

With:
C = costs;
y_m = output m.

The parameters estimated b_1 and b_2 here can be regarded as direct estimates of marginal cost of the corresponding services. In case C, y_1 and y_2 are measured in logarithms the estimated parameters b_1 and b_2 reflect cost elasticities with respect to services. In that case the sum of the parameters directly indicates the existence of (dis)economies of scale.

Note that this technique is quite different from stochastic frontier analysis. The mathematical representation (5-5) is nothing more or less than an general approximation of an arbitrary mathematical function at a local point. The parameters of equation (5-5) therefore need to be estimated for each observation separately. This is a cumbersome exercise. The parameters of (5-5) only makes sense in case that there are sufficient observations that can be marked as similar firms. This implies that with a sufficient number of observations the critique on SFA regarding the functional specification can be overcome. However, the way efficiency scores are derived from the estimates – as in SFA - stays worrisome. After local estimators have been established, the residuals for each observation can be calculated and similar techniques as SFA can be applied to these residuals to derive the efficiency scores. For further details see Fan et al. (1996) and for an extensive discussion of succeeding approaches, see Johnson & Kuosmanen (2015) in Ray et al. (2015).

Figure 5-7 and Figure 5-8 illustrate the technique by applying it to a cost function. The horizontal axis represent the amount of services y_1. The vertical corresponding total costs. In figure 5-7 the dots represent a set of observations and the blue curve a cost function estimated by applying SFA (in this case based on an exponential distribution).

In figure 5-8 for each observation the regression equation (5-5) is being applied. The regression line is then shifted to the most efficient firm in the sample (compare with aforementioned COLS). This is a rather simple way of doing. In Fan et al. (1996) a more sophisticated approach has been chosen, but we omit the details of these complex statistics. The green dot is the firm under investigation and the red dots represent all the other similar firms. The blue line represent the (shifted) regression line. The distance between the green dot (observation under investigation) and the blue line is a measure for inefficiency. From the first sample (on the left) it is obvious that the observation under investigation is very inefficient, the second and fourth subsample (at the right side of the picture) show firms with a modest inefficiency. The third subsample shows an efficient firm.

Figure 5-7 Frontier estimation based on exponential efficiency distribution

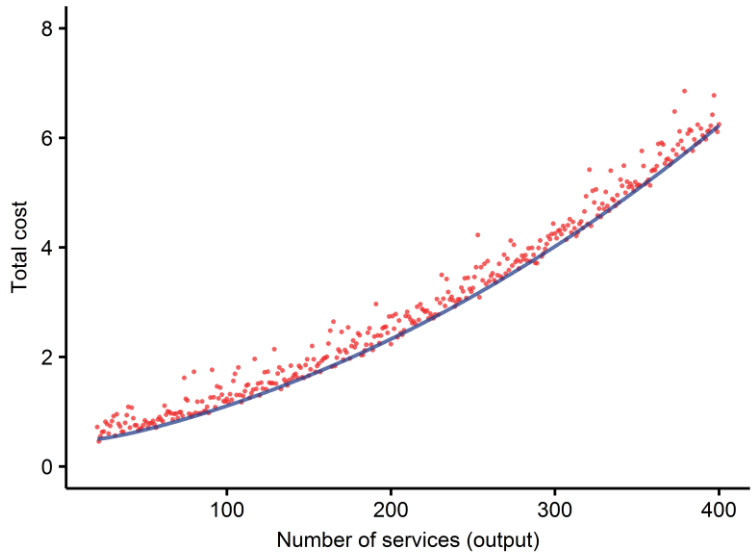
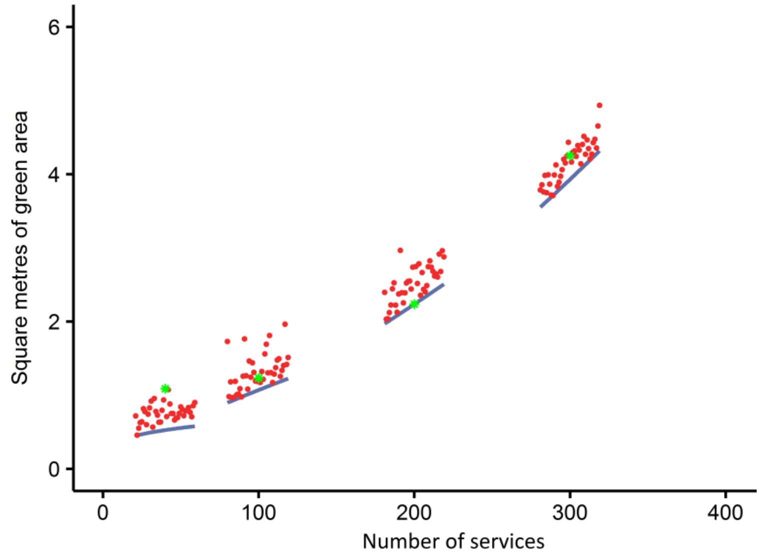


Figure 5-8 Frontier estimation based on local least squares method



The gradient of the blue lines also reveals an interesting issue. The blue line on the left is rather flat, implying that costs do not change very much as services increase. From this we can conclude that there are economies of scale. The blue lines get steeper as the firms under investigation produce more output. In the right upper corner of the picture we notice that costs grow faster than output. Here, diseconomies of scale prevail.

This technique heavily depends on what is defined as similar and how the weights are set. It is obvious that when the definition of similarity is tight or the weights are declining faster when the similarities wane, the estimates are accurate, but probably less precise. As said before, the technique particularly shows its strength in situations when there are plenty of observations available.

5.10 Efficiency scores and what's next?

So far we have presented different ways to calculate efficiency scores, but they are nothing more than numbers that reflect how much one firm is performing relatively better than another firm. This might be helpful in establishing parameters for funding systems or identifying bad performers for an inspection authority. However, the most important issue to address is how to perform better, in which case we need to get insights in the determinants that actually influence the efficiency. In Chapter 7 and 8 we discuss some of these determinants, but here we focus on how to measure the effect of these determinants on the efficiency. Roughly speaking there are three approaches:

1. Two step analysis;
2. DEA- bootstrapping;

3. One step analysis.

We consecutively discuss these three approaches.

5.10.1 Two-stage analysis

The most straightforward way to analyze the effects of the determinants is to conduct a multivariate regression analysis with the efficiency scores as the dependent variables and a number of determinants as explanatory variables. Assume that we derived the efficiency scores from a set of schools then we could estimate the effects of a number of managerial instruments on efficiency from:

$$Eff = a + b_1 Exp + b_2 Ttime + b_3 Class + ..$$

With:
Eff = efficiency score;
Exp = teacher experience;
$Ttime$ = teaching time;
$Class$ = class size.

The parameters a, b_1, b_2 and b_3 are reflecting the effects and can be estimated. This example has been completely elaborated for Dutch schools (see Blank, 2015). Nineteen determinants are being analyzed varying from teacher experience to the use of educational IT.

Since efficiency scores have a maximum value of one the dependent variable has a special distribution, not all the standard assumptions for regression analysis being fulfilled. In literature we therefore see techniques being used that takes this anomaly into account. Instead Tobit-analysis or maximum likelihood methods with various assumptions on the error structure of the model are used (see e.g. Kooreman, 1994).

Aside from this distributional problem there is another serious drawback of this approach. One of the problems that may arise is that the determinants may be correlated with the services produced. For instance it could be the case that large schools will have more financial opportunities to invest in IT. In that case the effects of school size established in the first stage will also absorb the effect of IT. In the second step no or a smaller effect for IT will be found than actually is the case. In statistical jargon we say that both estimated effects are biased. Some outstanding scholars in this field therefore state that "we hope to see no more two-stage SFA models" (Fried et al., 2008: p. 39).

5.10.2 DEA-bootstrapping

Bootstrapping is a technique used to approximate the sampling distribution of the dependent variable by simulating the data generating process (DGP). The objective of the bootstrap is to construct a confidence interval and a function of using bootstrap estimates of bias to correct for the bias of the DEA measures. This is crucial to construct a confidence intervals since each score is a function of other firms' performance. Interested readers are referred to the following video: https://www.youtube.com/watch?v=0hNQx9nagq4.

As Simar and Wilson (1998) explain, a distribution of the DEA score if used in a two-stage analysis regressing the score on a set of independent environmental variables is necessary since each DMU score is a function of other DMU's performance. Since the dependent variables are so intertwined, bootstrapping or rerunning the DEA methodology M number of times on a subset of the sample will lead to a confidence interval for each DMU, and therefore ready for the second stage analysis. See above for the video that explains bootstrapping in a tutorial.

An interesting application of the bootstrapping method can be found in Pilyavsky et al. (2006). In this paper, the authors compare Ukrainian hospitals by region hypothesizing that the hospitals operating in the West are more similar to the hospitals of western Europe and therefore more technologically advanced than the Eastern hospitals that are modeled after Soviet style hospitals.

Given the political differences between East and West, these authors also contend that there is more flexibility in the patient physician relationship that would explain the differences in the physician/patient ratio and the number of house visits (house calls). Conversely, in the East, bribes for treatment are more prevalent leading to a less consistent type of health care system.

Table 5-8 Major aspects of the research by Pilyavsky et al. (2006)

Sector	Ukrainian hospitals
Type of model	Output-based DEA with bootstrapping
Data	Ukrainian hospital data by communities 1997-2001
Production	Number of medical admissions, number of surgical admissions
Quality	Not included
Resources	Number of beds, nurses, physicians
Environmental factors	Eastern vs. Western region
Efficiency factors	Budget constraints, physician per population, percentage elderly, percentage surgical care, wage rates
Economies of scale/scope	Not included
Technological development	Not included
Efficiency scores	Hospitals in western Ukraine show greatest improvement in technical efficiency over time. Reduction in technical inefficiency over time depends on percentage elderly and percentage surgical care. More physicians per capita increase inefficiency.

5.10.3 One-step analysis

Considering the objections against the two-stage method it seems to make sense to include the determinants right away in the cost of production function and estimates the effects instantly. This is less straightforward as it seems, since we know that efficiency is bounded. A firm cannot perform better than the best practice, which is determined by the state of the art of the technology. In the past several techniques has been developed, in which the determinants are included in the error structure of the model and estimated with complicated maximum likelihood methods (see e.g. Battese & Coelli, 1995). We will not get any deeper into this techniques, but there is one more simple approach with a more intuitive appeal. We will discuss this here in more detail. The method is based on the so-called scaling property, which has not been explored widely in applications, but has been mentioned for quite a while (see e.g. Simar et al., 1994). Without discussing the special features of the scaling property It leads to a model specification that can easily be estimated with a (non-linear) least squares method and without any distributional assumptions about the efficiency component. To keep it simple, we start with the cost equation (5-5). Since this cost equation represents the frontier or best practice costs, the actual costs of a firm will be greater or equal to these costs. So actual costs equal frontier costs multiplied by a factor greater than one. We add this multiplication factor to the frontier costs. This yields:

$$C = (a + b_1 y_1 + b_2 y_2) \cdot e^{IF} \qquad (5\text{-}6)$$

With:
C = costs;
y_m = output m;
IF = inefficiency factor.

The inefficiency factor is a number greater or equal to zero (for instance 5%) and relates to a number of efficiency determinants. Since the exponential function always is positive, a good candidate is:

$$\boldsymbol{IF = e^{\{d_1 z_1 + d_2 z_2\}}} \qquad (5\text{-}7)$$

With:
z_1 = efficiency determinant 1;
z_2 = efficiency determinant 2;
d_1, d_2 parameters to be estimated.

Substituting (5-7) into (5-6) yields:

$$\boldsymbol{C = (a + b_1 y_1 + b_2 * y_2) \cdot e^{exp(d_1 z_1 + d_2, z_2.)}} \qquad (5\text{-}8)$$

Although equation (5-8) does not look like a standard regression equation (because the parameters d_1 and d_2 appear in the exponent), there is plenty of software available for estimating this type of equation.

Although the use of the scaling property in efficiency measurement has some major advantages, the approach has not been applied on a large scale (for further details see Alvarez et al., 2006; Schmidt, 2011; Wang & Schmidt, 2002).

5.11 Software

For SFA several software packages are marketed. The most extensive package probably is LIMDEP, developed by William H. Greene. Other econometric packages with standard SFA procedures include, SHAZAM and Stata. For those who want more flexibility we can recommend Eviews. In Eviews the standard methods can easily be programmed, but modifications of models can easily be added to the programming statements.

For DEA there now are several software packages available. The most common packages are: OpenSourceDEA, OnFront, PIM-DEA, EMS, DEA-solver, Frontier Analyst, WDEA, DEA frontier, MaxDEA, FEAR and DEAP. The choice of which software to use depends on individual choice, including convenience, applicability of the different models and flexibility of data input and output, familiarity et cetera. When researchers are familiar with the technique of linear programming, it will be sufficient to have access to software that contains the linear programming module, such as for example SAS. The proficient Excel user could easily process the DEA models in spreadsheets. We can also state that, as far as DEA software goes, it is more transparent and accessible than some of the econometric software. However, since Excel also includes Maximum Likelihood methods those with real interest and skill may be able to do some SFA as well.

6 Empirical research

6.1 Introduction

In this chapter, we set up an economic model to measure, to interpret, to estimate, and to present results.

6.2 The economic model

There is no such thing as *the* economic model or *the* productivity model. The setup of the research is determined by the definition of the policy problem. For example, if the policy problem statement is how best to reduce costs by increasing productivity, the question whether market forces have contributed to an increase in productivity is fundamentally different from the question whether *outsourcing* has affected the firm's productivity. This is especially true for questions that relate to a different view of a system's performances. When assessing the productivity of an agricultural firm, we can look at just the produced quantities of crops, or we can also check the contributions of environmental changes or the negative contribution to the environment of over-fertilizing (i.e., weak disposability). It is clear that the outcomes will differ substantially, because in this example there are several different policy problems that are being posed.. An individual farmer facing only market forces is concerned with prices and crop yields, namely, the cost function. The public, however, is also concerned with such externalities as pollution or environmental degradation due to irrigation. Therefore, "where you stand depends on where you sit." Similarly, a model for the strictly private sector is very different from a model in which the researcher wishes to include social goods.

Questions may relate to an existing or an historical situation, as well as to a virtual situation. An example of a virtual situation is the calculation of the effects of deregulation or of the introduction of a different financing system. By adjusting the model's setup, these types of questions can also be answered. The questions will often be directed at the present situation. If this is the case, it is important to match the question with the right model. In Chapter 3, we saw that economic behavior and economic constraints are important elements when choosing an economic model. Knowledge of institutions is therefore crucial when choosing a model. Do firms focus on the resources or on the services side? Are they limited to a certain budget or must they realize certain goals in terms of revenues? Researchers must make choices with regard to these issues and make the orientation of the model explicit.

6.3 Output, quality, and product prices

Services produced by firms in the public sector often cannot be easily measured. In the literature, we find a variety of product indicators per type of public sector. A number of considerations play a part in defining services:

- Focus of the study;
- Purposes of a service;
- Heterogeneity of service;
- Availability of indicators;
- Availability of observations.

Researchers must think carefully about what they intend to research as well as what data are available. It makes quite a difference whether the question focuses on efficiently supplying so-called intermediate products, end products, or final effects. For example, researching hospital productivity and efficiency is complex. Physicians who provide care in hospitals determine which diagnostics and treatments are available to patients. Hospital management is responsible for supplying, as efficiently as possible, laboratory research, numerous types of scans, other diagnostics, and nursing. For management it is vital to measure and to evaluate the production of these services. Management may also be interested in adjusting capacity in such a way that an optimal result in terms of the number of successfully treated patients is realized. In that case, the management also needs to ensure that specialists do not carry out unnecessary medical services or have patients stay in the hospital for longer than medically necessary. In the first case, the researcher will use the number of diagnostic tests and the number of nursing days as product indicators; in the second case, it will be the number of patients.

In the case of effects, the assessment will go even further, because the focus is on social goals, such as quality of life. In that case, it is more obvious to use the product indicators such as extra life years, corrected for quality.

For many services, there is a substantial heterogeneity in products. A hospital provides hundreds of laboratory tests and other forms of diagnostics using a variety of labor and capital inputs. Also the types of nursing care can differ substantially. For example, nursing the patients in an intensive care unit is much more demanding than nursing the patients on a general medical/surgical ward. In terms of end products, we usually speak of hundreds or thousands of different hospital treatments. Until 2011, the Dutch hospital system applied a system of product indicators that goes by the name Diagnosis Treatment Combinations. This system contained over 30,000 codes! In the US, there are over 900 diagnosis related-groups (DRG), so it has been suggested that analyses could be focused on one DRG to ensure homogeneity. A productivity analysis cannot handle such a number, which therefore requires a sensible reduction (see § 4.6).

Quality of service can also be seen as a form of heterogeneity. The quality of education can be represented by taking the difference between the number of failing and the number of successful students by means of the product indicators. Separate quality indicators are often added to the product indicators, such as the average exam score (of a school), the number of readmissions (of a hospital), or the number of power cuts (of an electricity provider). Quality of service can be challenging in decision-making discussions. Not only will the measurement encounter problems, but there is hardly any consensus about what must be included in quality. School management will probably greatly value high exam grades, whereas the Minister of Education, Culture, and Science and his or her policy makers value a low number of premature drop-outs.

For research purposes, one is often dependent on existing registration/data collection systems for securing the data required to answer the research question. The content of these registration systems limits the researchers' possibilities. Even if, for theoretical reasons, a certain set of indicators is to be preferred, researchers must instead work with a less ideal set of indicators.

The number of available observations determines the number of degrees of freedom in a research, i.e. the number of effects that can be estimated. It is no use trying to find out what the mutual relation among dozens of indicators is, when the researcher only has access to the data of just ten firms.

An important aspect of the public service is the nearly always absent product price. Many services have zero out of pocket prices for civilians (roads, street lights, and primary education) or have regulated prices that have little to do with a market valuation of a service. A tuition fee is the price you pay for education, but it is neither cost covering, nor a reflection of the individual appreciation of education. More and more market or quasi-market prices for public services are gradually becoming available, such as those for the liberalized segment of hospital care (segment of free prices). Product prices become especially important when the size and product mix are part of the discretionary authority of public firms. They can then also be assessed on strategic choices with regard to the size and mix of their services.

6.4 Resources and prices

6.4.1 Costs, quantities, and volumes

When employing resources, there are four values that we need to incorporate in order to build an economic model: costs, quantity, volume, and price. Costs refer to the nominal amounts in money that are spent on a certain resource. Quantity refers to the directly identifiable quantities, such as the number of employees or the number of kilowatt hours of power consumption. In addition to quantity, there is also the concept volume of resources. Volume takes into account the quality or the combination of the resource. The use of nurses is valued higher when there are many highly qualified or experienced nurses working in a firm, than when there are not. Quality and price are also directly related to volume, as wages will be based on qualifications (the human capital approach) and the supply of labor (the market approach). In principle, a researcher uses these volume units in analyses.

The most complicated group of resources to be made operational is that of capital goods. Machines, buildings, and transport are purchased at a certain point in time and are then used for many years. The purchase costs of capital goods are spread (depreciated) over a large number of years. The depreciation of capital goods is mostly used as an estimate of the use of capital in a year. The capital goods also represent a capacity that would yield a return on capital if it were available on the capital market. Potential interest revenues or payments (when the money is borrowed) are the second part of the capital costs. There are various systems to depreciate capital goods, and not every system is quite suitable to be used as an estimate. There are systems that base depreciations on the original purchase value of the capital goods, whereas others take into account the increased prices of the capital goods (so-called replacement value). For an extensive explanation of capital estimates ,we refer to (Lau, 2000).

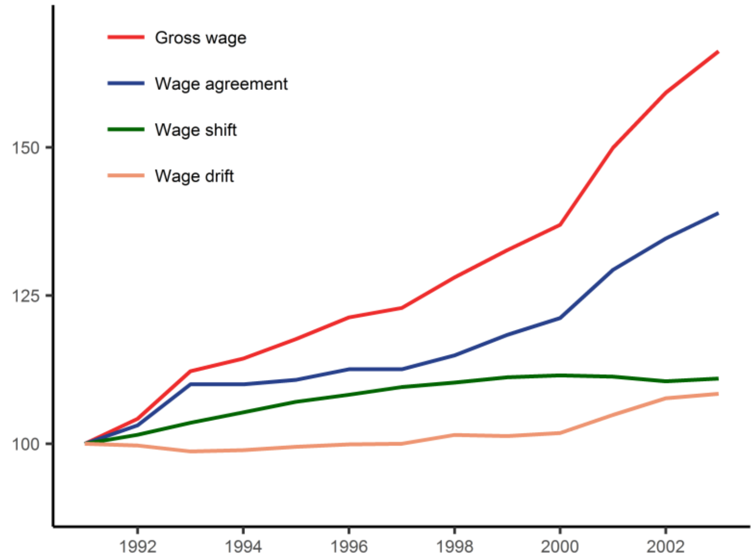
6.4.2 Prices of resources

Prices of resources refer to the price of a volume unit. Prices are often differentiated by region and time. Therefore, not only do prices change in the course of time, but there are also structural differences in prices among different regions. Examples of regional differences are housing (due to, for example differences in land prices) and employees (as a result of regional differences in labor supply). The volume of a resource is always the ratio of costs to regional resource price. Data are available on comparable cost of living among cities or regions; purchasing power parity (PPP) is also available to make comparisons among countries.

The measurement of prices is not always straightforward. National statistics agencies collect hundreds of different prices. A major distinction is made between consumer and producer prices. The former refers to the prices of final goods and services purchased by consumers, while the latter refers to the prices that producers pay for basic and intermediate goods and services (labor, materials, energy). As in the public sector the input of labor is dominant, including the right wage differences and growth are of great relevance. In general, labor contracts, wages, demographic shifts, and exogenous shocks to the market need to be included so that wage drift can be assessed. One can question whether these incidental wage rises should also be included in the price of labor, since they are partially commensurate with a higher quality and more experience of the staff concerned. That is why the incidental component is often further specified in a wage structure effect and wage drift. Moreover, the wage structure effect is linked to the quality of the staff, and wage drift to a better compensation due to a shortage of staff. In fact, wage drift is the only component that should be included in the price of staff. In Figure 6-1, the wage development for the various components in Dutch general hospitals between 1991 and 2003 is illustrated.

Public policy makers also focus on ensuring that there are sufficient providers in underserved areas. This is currently a significant issue in the US where effective health care reform is hindered by a shortage of primary care physicians.

Figure 6-1 Development of wages in Dutch general hospitals, 1991-2003



Source: (van Hulst, 2000; Vandermeulen et al., 1997; Visser et al., 2005)

Figure 6-1 shows the considerable wage development (almost 70%) in general hospitals in the Netherlands between 1991 and 2003. Two thirds of this can be recalculated to improvements of the collective labor agreements, and a third to improvements in the incidental atmosphere (wage shift + wage drift). More than half of the incidental component can be attributed to the wage shift effect. This type of wage drift versus wage development would be a useful analytical tool to apply to the public sector workers in the US who have recently been scrutinized by politically conservative state governors.

Prices are presented in index numbers. This means that the price in a certain year is set to 100 and the development is shown in relation to that price. Prices must be weighted by the relevance of the corresponding resources. There are various ways to do so. The most common price indices are the Laspeyres price index and the Paasche price index. The Laspeyres price index is based on weights of the reference year and the Paasche price index on the actual year. In case of two resource prices the indices are calculated as follows:

$$W_{t+1}^{LP} = \frac{w_{t+1}^{1}x_{t+1}^{1} + w_{t+1}^{2}x_{t+1}{}^{2}}{w_{t}^{1}x_{t}^{1} + w_{t}^{2}x_{t}{}^{2}} \qquad (6\text{-}1)$$

$$W_{t+1}^{PA} = \frac{w_{t+1}^{1}x_{t+1}^{1} + w_{t+1}^{2}x_{t+1}{}^{2}}{w_{t}^{1}x_{t}^{1} + w_{t}^{2}x_{t}{}^{2}} \qquad (6\text{-}2)$$

W_{t+1}^{LP} = Laspeyres price index;
W_{t+1}^{PA} = Paasche price index;

w_t^1 = price service 1 in period t;
w_t^2 = price service 2 in period t;
x_t^1 = quantity resource 1 in period t;
x_t^2 = quantity resource 2 in period t;

For example, the price of energy has an index of 115 in 2015 and the reference year is 2010, the price of energy increased by 15% in that period. The prices of the individual components in such a bundle must then be weighted and added up. Table 6-1 contains an elaboration of the price index number for energy.

Table 6-1 Determining price index for energy

	2010			*2015*		
	Price	Quantity	Costs	Price	Quantity	Costs
Gas	1	4,000	4,000	2	1,500	3,000
Electricity	2	6,000	12,000	3	5,000	15,000
Oil	2	3,000	6,000	1	7,000	7,000
Laspeyres		100			132	
Paasche		100			98	

Table 6-1 contains data on the prices and quantities of gas, electricity, and oil in 2010 and 2015. The year 2010 is the reference year. In 2015, gas and electricity have become more expensive, but oil has become cheaper. What does this mean in terms of energy prices? The overall energy costs in 2010 are €22,000. Weighting of the prices in 2015 with the 2010 changes this figure to €29,000 (=4000x2 + 6000x3 + 3000x1). This is a 32% increase in relation to 2010. The overall energy costs in 2010 are €25,000; weighting the prices in 2010 with the 2015 quantities results in €25,500. This means that the prices decreased by 2% between 2010 and 2015 (Paasche index of the general price level). The difference results from the fact that oil became cheaper between 2010 and 2015, and that the sold quantity in 2015 was considerably larger than in 2010. This method puts far more weight on the price decrease. Further, this approach can be readily applied to a cost-benefit analysis involving the assessment of changing costs due to energy versus sustainable, renewable energy sources.

The definition of prices in empirical productivity research can have a profound influence on the results. If you are familiar with any literature in this field, you will probably conclude that the attention for the definition of prices is in a poor state. The focus is much more on the definition of production, modeling, and the statistical methods. Therefore, it is not uncommon for researchers to use the *unit value*. The unit value is no more than the labor costs divided by the number of full-time jobs. In fact, any qualitative differences in employees may be attributed to the cost of employees. It will then seem as though some firms have more control of their technical process than others, whereas they employ more productive (but also more expensive) staff members. It is clear that the performances of the football team of Real Madrid are much better than those of Sparta-Rotterdam, because Real Madrid has the financial resources to contract the world's most talented players, and not because their business process has been organized much better. For US basketball fans, think of the Cleveland Cavaliers success after acquiring Lebron James.

The following example visualizes the influence and the importance of using the correct prices (Table 6-2). Assume that there are three firms, A, B, and C, each situated in a different region. Each firm produces 10 units using different staff employment, namely 15, 10, and 8 FTEs, respectively. The costs differ as well. The costs are 9 for firm A, 10 for B, and 16 for C. These variations are partially due to wage level differences between the regions; the regional prices are respectively 0.8, 1, and 1.5. When you take into account the differences in regional prices, the costs are 11.3, 10.0, and 10.7 respectively.

Table 6-2 Relation between units, volumes, prices, unit values and productivity

Data					
Firm	Services	FTEs	Costs	Regional price	Volume
A	10	15	9	0,8	11,3
B	10	10	10	1	10,0
C	10	8	16	1,5	10,7
Productivity					
Firm	Per FTE	Per euro costs	Per volume unit	Unit value	
A	67	111	89	60	
B	100	100	100	100	
C	125	63	94	200	

Calculating the productivity in terms of production per FTE, firm C has the highest score (= 100*10/8 =125). In terms of production per euro, this is firm A (= 111), and in terms of services per volume unit, firm B (= 100) prevails. Firm A has the cheapest employees (unit value is the lowest of the firms). Firm A operates in a less expensive region, and also pays its employees low wages. This firm mostly employs semi-skilled and unskilled staff, which is also reflected in a low productivity per FTE. On the other hand, firm C has a very high labor productivity, but also pays the highest price. The unit value equals 200.

Researchers must be aware of these differences and deal with them accordingly. In most situations, it would seem obvious to use the productivity in terms of output per volume unit. In this approach, there is a correction for the effect of prices established outside the firms. When a firm, for example, pays salaries that are above the prevailing market-rate salaries, these will be counted as inefficiency (the volume increases).

6.5 Environmental factors

Environmental factors refer to circumstances that affect influence productivity but either cannot, be influenced by decision makers or the decision maker has limited influence. A severe winter automatically leads to higher energy use, and a flu epidemic will lead to more absenteeism through illness. In both cases, there is a negative effect on productivity over which the management of a firm has little or no influence. However, effective leadership could anticipate environmental factors. For example, offering home based computers for employees to work at home in the face of a snow-storm or arrange for safe transportation for essential personnel. The impact of a flu epidemic could be mitigated by a firm offering vaccines in advance of flu season. These scenarios and social are different from other external factors.

Knowledge about a sector's institutions and organizations is important in this respect. In education, the local government is largely responsible for school buildings. When municipalities have old schools buildings or their schools are in temporary buildings, the schools will face high energy and maintenance bills. When schools are then assessed on their productivity in a benchmark, it is only fair to correct for these environmental differences.

In Chapter 3, integral models were discussed that easily allow for a correction. By adding variables about the environment to the model and collecting data about them, the effects of the environment can easily be empirically determined. In the previously mentioned examples, the average temperature and the number of flu reports at the health clinic or physicians' offices could be collected. In the example of school buildings, the average age of the building and the number of square meters of building surface of temporary or semi-permanent buildings can be included in the model.

6.6 Environmental factors and educational performances

Afonso & St. Aubyn (2006) research the efficiency of secondary education in various OECD countries (Organisation for Economic Co-operation and Development). Differences in efficiency are explained by the influence of environmental factors.

Table 6-3 Major aspects of the research by Afonso and Aubyn (2006)

Sector	Secondary education
Type of model	Output-oriented DEA with 2-stage analysis
Data	Cross section of 25 countries
Production	PISA-test scores (4 dimensions)
Quality	Included in production
Resources	Teaching staff, instruction time
Environmental factors	GNP per capita, educational level of parents
Efficiency factors	Not included
Economies of scale/scope	Not included
Technological development	Not included
Efficiency scores	After correction, between 0% and 25% (average 12%)

Alfonso and St. Aubyn researched the efficiency using a so-called 2-stage DEA. The first stage consists of determining the efficiency scores. The second stage consists of measuring the influence of environmental factors on the efficiency scores. The method was applied to a data-set of 25 countries. The observation unit was therefore a country and not a school. For each country, there are data on the performance of 15-year-old students in the international PISA test, which tests certain abilities, such as arithmetic and reading. The resources are the number of teachers per one hundred students and the average number of classes per year per student. The observation period for the PISA scores is 2003; for the use of resources this is the sum of 2001 and 2002. The average inefficiency is about 12%; the scores vary between 0% and 45%. This does not necessarily mean that a country with a low score has organized its education in a poor fashion. Environmental factors can have a profound influence on the calculated inefficiencies.

In the second stage, the researchers measured the influence of environmental factors, such as the economic development of a country (GNP per capita) and the level of training of the parents. By correcting the original scores for the influence of the environmental factors, we get an entirely different picture. The scores then vary between 0% and 25%. The dispersion is smaller, whereas the average inefficiency increases to 13%. The order of rank for efficiency also changes drastically. A country such as Norway is ranked 16th on the original list, but drops to 25th on the corrected list. Portugal then climbs from 20th on the original list to place, to 1st in the corrected list. This research is a refreshing example of the cautiousness with which the results of efficiency research must be interpreted. Inefficiency is often associated with a poor organization or management and often serves as a reason for implementing policy changes but including circumstances outside the influence of the responsible actors adds information toward solving the policy problem.

6.7 Evaluation of government policy: financing, management supervising authorities

A great deal of research on the productivity of public services results directly from policy issues. Questions concerning what will be dealt with include:

- Permits for supplying a public service;
- Legislation about quality and safety regulations;
- Financing public services;
- Capacity growth and planning;
- Ownership;
- Profit;
- Contract agreements;
- Reporting and responsibilities;
- Registration obligations;
- Free pricing/regulated prices;
- Degree of market forces/competition;
- Inspection and supervision.

This is only a fraction of an range of government instruments that may affect productivity or efficiency. Here, we discuss the effects of each of these instruments not in detail, but globally, by summarizing them in two groups: a group with variables in the discretionary space of firms, and a group of variables concerning the stimuli structure in the system.

The variables related to the discretionary space refer to the freedom of management to set up the production process or determine the quality of the offered services. Examples are the license and quality demands, such as the obligation of secondary education schools to plan a minimum number of teaching hours per year. Such demands limit the space of school management and boards.

The variables concerning the stimuli structure measure the external pressure to perform well. This is mostly expressed in such aspects as ownership, profit or non-profit, free pricing, and the strength of market forces.

The line between the groups is not always apparent. Housing requirements belong to the group relating to the discretionary space, as well as to the group relating to the stimuli structure. The reason is that housing requirements can be seen as an admission threshold that can influence the degree of competition in a market.

In empirical research the policy variables often have a typical character such as a yes/no situation. In that case, the effect of this variable cannot be determined, because it is not possible to distinguish or compare firms with or without profits. The effect can sometimes be measured when there is a change in policy, causing changes to be measured from one year to another, thus indicating the effect of the policy change. The restriction of this approach is that there can be more changes in that same year (e.g., technical changes) that also have an effect that cannot be distinguished from the policy effect.

A number of variables mentioned in the list are hard to quantify, because they have a more qualitative nature. It is difficult to assign a value to contract agreements, documentation, and responsibility. In these cases, the researchers will develop a number of typologies, whereby each typology reflects a number of contracts containing agreements. Contracts can be divided into open-end contracts, performance contracts, and mixed contracts. Open-end contracts feature a settlement for each service. In performance contracts, the amount and the quality that must be provided are documented. Mixed contracts comprise aspects of open end and performance contracts.

Yes/no and qualitative variables can be included in empirical research as so-called 0-1 variables or *dummies*. The estimated parameter can then indicate the effect of the use of a performance contract instead of an open-end contract. Some variables, such as the degree of competition, can easily be used for a direct measurement. For this purpose, various indicators have been developed, such as the Herfindahl index (HHI), which is the sum of the square of all market shares. The minimal value of the HHI is 1/N (N = number of firms) when all firms have an equal market share, and the maximum value is 1 when a firm serves the entire market.

6.8 Evaluation management: process, control and strategy

The management's influence primarily manifests itself in setting up the production process and making strategic choices. The process especially focuses on setting up the output of goods and services, and on answering such questions as which production resources are used, how these are logistically linked, and which maintenance and replacement schedules are utilized. These are major questions in the production process. In sectors with a relatively high staff employment, there are similar questions, such as about working schedules, labor conditions and human resource management. This is why research includes such variables as the part-time work factor, the age groups and the educational level of staff, the prevention of absence policy, the age and the user intensity of equipment, maintenance and check-up schedules, outsourcing of services, purchase agreements, and stock control. Policies as paid childcare leave or leave to care for elder parent care also needs anticipation and utilizing plans how to fulfill the task usually produced by the employee taking leave is important for successful management.

Control relates to the way the entire process is managed and controlled. This concerns management layers, the distribution of responsibilities among management layers, the supervision within a firm, and procedures and protocols. Examples of control variables are: the size and quality of the supervisory board, the board of directors, and the management of trading firm departments; the authority of each management layer to make decisions about what financial obligations are to be committed to; who decides about hiring employees, and the degree of participation of professionals (specialists on a hospital board), the number of formalized board meetings (for example, the supervisory board), bonus arrangements, and the presence of certification requirements.

Strategy is about the long-term goals of firms. Strategy especially refers to the orientation of certain market niches, and the development of new products and collaborations. Examples are the scale and the diversity of the output, forms of collaboration (chains, networks), contract agreements with employees, and market concentration. An interesting example is the cooperation between municipalities for local tax levying (Niaounakis & Blank, 2015, 2017).

7 Policy analysis

7.1 Introduction

The focus of this chapter is a number of policy themes in which productivity research can play an important part. We write about the following themes of the policy analysis process: capacity planning (accessibility and availability), financing/defrayment, market forces (prices, entry, and withdrawal) and innovations (technical and social).

7.2 Capacity planning

Productivity research can be crucial in capacity planning. Capacity planning refers to determining the level and composition of the supply of services and the spatial diffusion of services.

It is clear that this determination can only be made when the criteria that a certain configuration must meet are known. Must we meet a cost criterion that is as low as possible, must the travel distance for users be minimized, or must we meet certain quality criteria? There are also numerous constraints. Perhaps the government does not want to dictate where certain facilities must be located. Two examples of capacity planning are given here. The first relates to education. The freedom of choice of education is a social good in the Netherlands, and when parents are free to choose their children's education, it is pointless to have the government dictate the variety of schools. In such a case, the government uses global panels, such as foundation and discontinuance regulations. Moreover, these depend on local situations, such as the density of the population and the existing schools. A second example is the capacity of intensive care units (ICUs) in Florida in the case of a simulated hurricane in Miami requiring that patients need to be moved to other hospitals in Florida. The criteria included available capacity as well as the medical capability needed to treat the specific types of patients – cardiac care, neonatal care, pediatric care, and surgical medical care (Valdmanis et al., 2010).

Table 7-1 Major aspects of the research by Valdmanis et al. (2010)

Sector	Hospitals and intensive care units in Florida (US)
Type of model	DEA Plant Capacity Model
Data	Hospitals and ICUs 2005; Florida Care Agency for Health Care Administration, medical hospital cost report, Solucient, Inc (case-mix indicators)
Production	Available capacity medical/surgical ICU patients, cardiac care unit patients, neonatal unit patients, pediatric intensive care unit patients
Quality	Not included
Resources	Beds (per unit), medical staff (per unit)
Environmental factors	Closure of hospitals in Miami
Efficiency factors	Location
Economies of scale/scope	Not Included
Technological development	Not Included
Efficiency scores	Capacity rates ranges from 0.83 to 1.00 17% to 0% excess capacity

Therefore, a number of considerations are important when planning capacity. The cost aspect mostly refers to the scale and diversity of the supply. Knowledge about economies of scope and scale are necessary to start from some sort of optimal firm when planning. At a given demand for a facility, the number of firms and their degree of specialization or diversification follow automatically from the optimum. From this optimal number of firms, we can deduce what the consequences are for the quality and other external costs. A large fire station may be optimal from a cost point of view, but not from a safety point of view. Fewer fire stations will increase the travel distance, increasing the chance of property damage and/or loss of life.

External costs are costs that must be supported elsewhere. A decrease in the number of schools will lead to higher costs for students and parents in terms of travel distance and transport costs. Here, we may speak of a trade-off. An interesting example of such a trade-off and its optimization can be found in Blank (1993). In an analysis he shows what the consequences are, in terms of social costs, when adjustments are made in the supply of primary schools in the Netherlands. In the US, vouchers using the market for parents to select schools for their children including private and religious schools are predicted to hinder funding for US public schools. See for example an analysis done by the Brookings Institution: https://www.brookings.edu/research/on-negative-effects-of-vouchers/

It is not always the central government that sets up rules for capacity; the task is sometimes decentralized. In primary and secondary education, the local government is largely responsible for investments in school buildings. For many facilities the responsibility has been entirely decentralized toward the firm itself. The firms then receive an integral budget for developing capacities. They determine where, how much, and what kind of capacity is available.

7.3 Financing

There are various systems for financing public services. They vary from entirely input aimed, to a 'free market' financing. Departments, for example, are input aimed meaning that the number of employees and the extent to spend on material supplies are fixed. There is little room for a department to reallocate the resources. The other end of the spectrum consists of systems that demand that firms must generate as much income as they need to cover their costs. The most striking example is the part of Dutch healthcare that is covered by the Care Insurance Act. Healthcare firms supply services that are invoiced to the patient or his or her insurer. The prices for these services are still largely regulated. How the money is utilized internally is the business of the healthcare firm. Naturally, the healthcare firm must meet quality demands. The trend in recent years is that the prices for care services are also becoming a matter for the market. About 70% of the revenue of hospital care is already under this free prices regime (the so-called B-segment). It should be clear that each financing system has its own positive or negative stimuli for a firm's productivity. Firms with few possibilities to allocate their resources may not be held accountable for any allocative inefficiencies. In that case, the government is responsible. In some situations, firms are even encouraged to display inefficient behavior, for example when they are allocated a fixed budget and cannot retain any of the excess.

An important part of the financing system determines the financial parameters. These parameters indicate how large a budget must be in a certain situation or which tariff/tax is used for the supply of a service. It is therefore not easy to determine the accurate parameters for multiple services. With the help of productivity research, cost prices for the various services can be determined when production is efficient. There have been complaints by private insurance firms that they cannot continue to participate in the mandated exchange market in the US Affordable Care Act. However, as in all for-profit firms who answer to shareholders, Aetna, a for-profit health insurance firm claimed such losses. However, a judge found that Aetna's claims were false (http://www.latimes.com/business/hiltzik/la-fi-hiltzik-aetna-obamacare-20170123-story.html.) Therefore, any financing decision based on productivity must also examine underlying motives, particularly with contracting with for-profit firms for the provision of social goods/services.

Productivity analyses offer a fine evaluation instrument to assess a change in financing systems ex post. The instrument can also be used for an ex ante evaluation. With knowledge of the production function, it is possible to calculate the optimal allocation of resources when there is a change in the financing system. An interesting example can be found in Grosskopf et al. (2000). They discuss a model to simulate the potential gains in school efficiency in Texas school districts from reducing resource regulations.

7.4 Market forces

An often-heard opinion is that market forces lead to more competition and to higher levels of quality, to greater efficiency, and more innovations. Productivity gains from market forces in every possible way. Productivity analyses should be able to verify this hypothesis. By including the market proportions as an explanatory element or environmental aspect in the analyses, it should be possible to test these assumptions.

Market shares are often expressed in concentration measurements. Section 4.6 elaborated on a number of concentration measurements, such as the Herfindahl index and the Gini coefficient. These may also be applied in this case. Instead of shares of services within a firm, we now talk about market shares. A problem for these kinds of researches is defining the geographical regions in which the concentration must be calculated.

The literature on this item is impressive. Much research has been carried out in the United States, especially in the field of healthcare. Interesting overviews can be found in, for example, Mutter & Rosko (2008) and Sari (2008). A great deal of research has also been conducted into network sectors. The numerous liberalizations of the energy and public transport markets in the last two decades have contributed to this (Barnum et al., 2007; Button & Costa, 1999; Cullmann et al., 2008; Plagnet, 2006).

Here, we present an example from Bates et al. (2006), who conducted empirical research into the influence of market structure aspects on the technical efficiency of hospitals in various metropolitan areas in the United States. The research consisted of two phases. In the first phase, the researchers calculated the efficiency scores of the various areas. In the second phase, they related the efficiency scores to market structure aspects by means of a multivariate regression analysis. The aspects of the market structure entailed the degree of competition between hospitals, the activities of the Health Maintenance Organizations (HMOs), the degree of concentration of private insurers, and the presence of a so-called Certificate of Need Law.

Table 7-2 Major aspects of the research by Bates et al. (2006)

Sector	Hospitals
Type of model	Input-oriented DEA with 2-stage analysis
Data	Cross section of 306 metropolitan areas
Production	Treatment days, operations, visits to policlinic, births
Quality	Not included
Resources	Specialists, nursing staff, other staff, material expenditures, beds
Environmental factors	Number of hospitals per 100,000 citizens, share of Medicare patients, degree of penetration of HMO-s at the state level, market share of the top three insurance firms at the state level, certification law effective, presence of academic hospitals
Efficiency factors	Not included
Economies of scale/scope	Constant returns to scale assumed
Technological development	Not included
Efficiency scores	Between 0.59 and 1 (average 0.89)

The efficiency scores between the different areas vary substantially (50% to 100%). The most important explanatory market factors are the share of hospitals with an HMO contract and the market share of the three major insurers in the area. Both factors have a positive effect on efficiency. A negative effect was found in hospitals with a training function.

In Section 6.2 we pointed out the importance of calculating the "right" financial parameters. These are based on efficient cost prices. The same approach can be applied in view of monitoring. Information about market prices can be confronted with efficient cost prices to assess whether firms are abusing their market power to increase the prices.

In the example of OECD total factor productivity, only one output was specified. As we have seen, however, there are methodologies that can handle multiple outputs. Khan (2005) studied the total factor productivity in Pakistan (see Table 2-2) and included other finessed measures, such as human capital development (education) and economic openness (imports/exports). Given more detailed data-sets, environmental factors such as health status of the population, environmental degradation, and political freedom can also be included as outputs in measuring not only economic but also social well-being.

7.5 Innovation

Innovations can make an important contribution to productivity growth. Many factors play a part when developing new technology and its spread. These factors include organizations' tendency toward change, market organization, financial stimuli, institutional impediments, and the costs of implementing and using this new technology.

Insight into the effect of these factors is important for the development of the policy for stimulating innovations. A crucial factor for implementing an innovation is the expected productivity growth. In healthcare, for example, productivity growth is mostly a matter of quality growth. Specialists often wish to use the latest techniques in order to better diagnose and treat their patients, and the suppliers of new techniques are eager to sell them. Specialists wield the power to implement new techniques, but do not need to pay for them. In such a situation, a sensible cost and return consideration is missing.

Clarifying the productivity growth of some innovations is therefore very important when innovations spread. Innovation monitors can play a significant part in this. The effect of an innovation will be visible only after a few years, because sufficient observations must be made of firms that have implemented the innovation.

Furthermore, it must not be forgotten that the front-runners in innovations are firms that strongly focus on improving management or have sufficient financial resources. The productivity growth could then be wrongfully attributed to certain innovations.

The opposite may also be true. Where there is a strong focus on innovations, counter forces will occur. These counter forces have an interest in the existing situation or arise from an intrinsic resistance to change. Innovations sometimes come to a premature end or involve high costs. In the latter case, there is a productivity decrease. This is known as the innovation paradox (Deloitte Research, 2004). In relation to developments in the IT sector, Solow (1987) said of what he called the productivity paradox: *'You can see the computer age everywhere but in the productivity statistics'*.

In the previous paragraphs, we clarified that the contribution of innovations to productivity growth is hard to determine. An improvement will usually not become visible until after a few years. Processing certain delays in the measurement models is an option, but it also increases the models' complexity. It would seem more obvious to develop typologies for the innovation power of firms and to determine the connection between typologies and productivity. Typologies may be differentiated into various kinds. An example of such an approach can be found in Section 8.5 (innovations in hospitals) or in Haelermans & Blank (2012).

8 Management and strategy development

8.1 Individual benchmarks

Individual benchmarks of firms may contain very valuable information that can be used to improve management or to help with strategy development. For management, the information can be about subjects such as control and internal organization, staff policy (human resource management), outsourcing, and purchasing. In strategy development, the emphasis is on developing new products, entering into alliances, acquiring firms, or merging with other firms. Even though they appear to be primarily private market issues, these managerial issues can also be applied to the public sector, especially if lessons can be learned to lead to more efficiency in the social service production or mitigating negative externalities such as reducing fossil fuels and their attendant pollution for renewable energy sources.. This chapter will delve further into these various facets. Each section is related in one way or another to productivity. What we present here is a global insight into a number of essential aspects of management and corporate strategy.

8.2 Management

8.2.1 Control and internal organization

The internal organization and control (which are collectively known as corporate governance) are important determinants of productivity/efficiency. Matters such as the number of layers in an organization, the responsibilities and authorities of each layer, the internal registration systems, the internal control mechanisms, internal stimuli, the degree of participation of the employees, and the degree of transparency are dealt with. This is but a fraction of the range of aspects of the internal organization. Further refinement is possible just within the aspect of authority and responsibility. Authorities may relate to the financial aspects of the business (purchase, sales, the maximum amount there is authority to sign), as well as to the staff policy (recruitment, dismissal, labor conditions) and investment decisions.

8.2.2 Staff policy

Staff policy- or human resource management (HRM) is another important determinant of productivity, particularly in the public sector, due to its high labor intensity. HRM has a profound influence on issues such as staff turnover, absenteeism through illness, and work satisfaction. These can have a substantial impact on a firm's productivity. In recent research, physicians who were formal employees of hospitals resulted in better efficiency and quality of care of services provided. The theoretical underpinning is that physicians who are formal employees have more at stake in the success of the hospital (Goes & Zhan, 1995).

A firm's management has a number of instruments at its disposal for developing an effective staff policy. Think of recruiting and selecting employees, working conditions, assessment and evaluation talks, career coaching, training facilities, labor conditions, and absence prevention and guidance. The last-mentioned group includes workload assessment, absence reports, procedures related to absence guidance, and the content of contracts with safety, health, and welfare services. It has also been shown that working between 34-37 hours per week had higher productivity per worker than workers who put in long hours such as over 40 hours per week (Goes & Zhan, 1995).

Absenteeism has an immediate impact on productivity. An absence rate of 10% results in a productivity decrease of 6% when the labor intensity is 60%. The same decreases come about in the event of staff turnover. A substantial level of turnover leads to a substantial loss of productivity, not only because of the new vacancy, but also because a new employee must be sought, selected, and trained, and that entails costs. Suppose that 10% of the desired staff composition is continuously open, then (at a labor intensity of 60%) the productivity decrease will be around 6%.

8.2.3 Outsourcing/contracting

Outsourcing means that a firm has another firm execute part of the business process, usually supporting services, such as administrative tasks, IT, and housekeeping.

Outsourcing has numerous benefits. The firm does not need to worry about hiring qualified staff or business resources for this part of its process. The contracted firm is a specialized firm that not only supplies high quality, but also does so more efficiently because of the scale and specialization. The outsourcing firm often has fewer expenses than is the case when it executes these tasks itself. Moreover, the firm transfers part of the financial risks to the contracted firm, especially when the costs are determined based on a declaration of hours. A contract for a fixed period or with a stipulated term of notice also limits the risk. Naturally, outsourcing also has negative aspects. For example, because of its poor knowledge about the outsourced business activity, the outsourcer might pay too high a price or be offered inferior quality by the contracted firm. The outsourcer does not have any insight into the process at the contractor. A bankruptcy or a sudden stagnating supply of semi-finished products can suddenly create great difficulties for the outsourcer.

Outsourcing is used more frequently in some parts of the public sector than in other parts. It is common especially in healthcare. Hospitals outsource not only housekeeping or IT services, but also the medical treatment, the nursing, and parts of the infrastructure. Most specialists are not employed by hospitals, but work for independent ventures which has been shown to decrease efficiency (Al-Amin et al., 2017).

As far as nursing is concerned, employment agencies for healthcare and freelance employers are more frequently contracted than formally employed by hospitals. Project developers, banks, and insurers manifest themselves more explicitly when it comes to hospital buildings and the medical infrastructure. The full control and maintenance of buildings and the medical inventory are then outsourced to other parties. Parts of the primary process are also outsourced in public services where you might not expect it, such as the police. It is common knowledge that municipalities sometimes hire private security services to patrol burglary- sensitive neighborhoods or use citizen based neighborhood watches.

8.2.4 Purchase

The purchase of materials such as energy, phone services, medicines, and fuel for service cars can also affect a firm’s productivity. The presumed effect is related mostly the purchase price, and much less to the technical part of the production process (even if the effect is present). Price benefits occur when there is bulk purchase realized by joint purchase.

Apart from the price benefit it is also possible to benefit from a firm’s central point of purchase. A central point of purchase can realize economies of scale that cannot be achieved by decentralized purchase by departments. Within the organization, fewer employees need to handle purchase tasks in this situation. Another possible effect on productivity has to do with keeping stock. However, in that case one has to deal with the cost of keeping and controlling stock.

8.2.5 Occupancy rate

The occupancy rate reflects what part of the available production capacity is actually used for supplying goods and services. Examples are the percentage of occupied beds in a hospital, the percentage of hours that operating theatres are used, and the percentage of hours that a classroom is used.

To a large extent, the occupancy rate is determined by planning and schedules. Managers will try to maximize certain goals by optimizing the use of the available capacity. Accounting for capacity relies on solving logistical issues, for which you also have to take into account numerous constraints and many uncertainties. It is obvious that it is far more difficult for a hospital to set up a static working schedule, due to a high degree of unplanned demand as compared to the planning processes for a school. But in schools, the planners also face uncertainties concerning the development of the number of students and the sudden illness or resignations of teachers.

The occupancy rate can also profoundly impact the service. A high occupancy rate may, at first sight, appear to indicate high efficiency. In practice, however, a high occupancy rate may be at the expense of the service's quality. Especially in emergency services, such as the police, fire department, and ambulance transport a high occupancy rate often leads to not having these services immediately available when unexpected calamities occur.

An example was in Boston April 15, 2013 where all elective surgeries were postponed to permit emergency surgeries for victims of the Boston Marathon bombing.

8.2.6 Lease or purchase of buildings and fixed assets

When employing fixed assets, a firm must make different choices. Fixed assets are resources that have a life span of more than one year. Examples include the car fleet, equipment, IT resources, and business premises. The firm can buy these new and then decide when replacement or modernization is necessary. There are other options, such as lease or rent. In fact, these have the same conditions as outsourcing.

In the public sector, it often occurs that housing is not an immediate part of the discretionary authorities of the firm. In healthcare, capacity has existed for years and there have been lengthy and complicated planning procedures for the expansion of capacity. A considerable part of education is still under the purview of the local governments that handle the infrastructure of schools. The thought behind this is that public firms must not be burdened with parts of the management that involve a risk element. It is clear that municipalities can shift school capacity more easily than schools can.

Using fixed assets influences not only the use of resources, but also the use of other resources. Housing a school in an old and poorly insulated building means that energy and maintenance costs will increase considerably. Firms that have both a variable and a fixed employment of resources can make efficient choices, unlike firms that are regulated by the government and have to deal with a fixed use of resources. They latter may be in a situation of allocative inefficiency over which local decisions makers have no influence.

8.3 Corporate strategy

8.3.1 Scale

The scale or the size of the firm is of profound importance to the output and supply side. At the output side, the scale effects, as discussed in Subsection 3.2.3, are important. By increasing the scale, cost savings can be realized. This often applies only up to a certain level. When this level is exceeded, diseconomies of scale occur. Small firms often aim for an increase in scale for cost reasons. As this is not applicable to large firms, there must be other reasons why larger firms exist.

One of these reasons is market power. With an increase in scale, the market share grows. This larger market share can be translated into market power, which enables firms to largely determine at what price and quality the services are offered. This applies not only to the free market, but also to a strongly regulated market. A large market share often leads to practicing political bargaining power.

Market power can also be exercised on the labor market or purchase market (see also § 7.4). The regional labor market for some professionals, such as nurses, can only choose from among a few employers. This worsens their negotiating position regarding labor conditions. This influence is usually limited, because employees can move to other regions or similar sectors, if they find themselves in an oligopsony.

The scale aspect is also important for the development of individual strategies. A larger scale of the firm usually also means a higher salary scale for managers. A larger firm also leads to more respectability and status for the managers. These are institutionalized stimuli that do not contribute to a more efficient management or an improvement of the service in any way.

8.3.2 Diversity

Diversity refers to the combination of the number of services that are offered by a firm. The management of a firm is often faced with having to decide whether the firm should specialize in a certain service, a number of specific services, or a gamut of different services. Just as in the matter of optimal scale, there are considerations that concern the service supply side. We identified diversification in Subsection 3.2.3. We speak of economies of scope when the combined production of two services results in cost reduction with regard to the individual production of the two services. For instance, the supply of gas, electricity, and water by the same firm is profitable, because the firm only needs to send one invoice to each customer and only needs to process one payment for three services.

There is a similar argument on the supply side. More diversity will lead to a larger market share and therefore to more market power (think of the food industry). Diversity is sometimes also a means to further increase scale. In a saturated and concentrated market, a further increase in scale for a product is only possible by means of diversification or diversifying into other markets. This means that the firm becomes active in other markets.

Diversification also means spreading of risks. A firm becomes less sensitive to strong movements in the demand for a certain product. Because of a specific technical change a competitor could realize a large market share at the expense of other suppliers. Specialized firms could suffer most from this. In fact, a firm spreads the risks of suboptimal capacity usage, as discussed in Subsection 8.2.5.

8.3.3 Collaboration

Economies of scale and scope can sometimes also be achieved via collaboration. Collaboration between firms can be realized in various ways. The most rigorous form is when firms are combined in a holding company or under a combined board. The umbrella board layer then has a number of authorities that can be exerted upon the various parts of the holding. This makes it possible for the supply of products and services to be divided over the different parts, or for a number of facilities (such as staff administration and IT) to be shared by the different parts. Numerous modalities are possible. Education in both the Netherlands and the USA has a strong tendency toward large school boards (Onderwijsraad, 2008). These school boards sometimes control dozens of schools and ten thousands of students. The staff are employed by the school board and the school board supports the management of the schools in various ways. The schools are especially responsible for the primary process, namely the content and process of education. Collaboration among various public sectors is crucial but may be problematic due to “turf” issues.

8.4 Innovations

Another important aspect of the corporate strategy concerns the development and implementation of new techniques in the firm. In the broadest sense, this innovation strategy relates not only to new equipment, but also to setting up technical processes and the functioning of the relations between different stakeholders (so-called social innovations). Logistical issues and those relating to the functioning of networks and chains are therefore also expressly included.

Innovation is about strategic decisions regarding investment in research and development and in applying new techniques. The firm’s management must consider the costs of developing and gathering new knowledge, the costs of implementation and the benefits of productivity growth, and the development of new markets and clients resulting in the positive externality of job creation and general well-being of individuals. These considerations mostly take place in a situation with many uncertainties, because little information about the effects of new techniques is available.

8.5 Effect of innovations on hospitals

Blank & van Hulst (2009) researched the effect of innovations on the productivity of hospitals in the Netherlands. Table 8-1 presents the major aspects of their research.

The research distinguished 63 innovations. For each hospital, there are data on whether an innovation has been implemented and, if so when this took place. The 63 innovations were grouped into 7 clusters (e.g. medical technology and IT). The innovations in a cluster were included in the analyses as a determinant. The research shows that some innovations influence productivity, while others do not. A significant positive influence can be found in the care and logistics chain. Innovations in multidisciplinary diagnostics and outpatient care negatively influence productivity.

Table 8-1 Major aspects of the research by Blank and Van Hulst (2009)

Sector	Dutch hospitals
Type of model	Cost function
Data	Annual survey of hospitals 1995-2002
Production	Admission according to specialty
Quality	None
Resources	Management, nursing staff, supporting staff, material, capital
Environmental factors	None
Efficiency factors	Number of innovations, divided over 7 clusters
Economies of scale/scope	Not included
Technological development	Approximately 1% per year
Efficiency scores	Not included

8.6 Strategy and productivity

There is an extensive literature on scale effects. This is not the case for diversification effects. Various studies discuss the aspect of collaborations. For example, in the United States there is a large diversity of hospital configurations. These have been elaborately studied by various researchers (Alam & Granderson, 2008; Bazzoli, 2008). However, information regarding the ultimate effects of the innovations is often scarce. An innovation must first be implemented before research can determine its results. And this is the exact same period during which the managers of other firms must decide whether to adopt the innovation. The moment that research results are available they are no longer relevant to managers. However, research that comments on the determinants of the spread of innovations is available (Blank & van Hulst, 2005). This knowledge may also be relevant to managers by providing an insight into the circumstances under which innovations are rapidly implemented.

9 Concluding comments

In this book, we have attempted to bridge the gap between economic science and practice with an eye toward assessing the productivity and efficiency of the public sector. We have explained theoretical issues in a comprehensible manner. With the help of practical examples, we applied both theory and modeling to sectors of the public/social sector, including health, education, environment, police services, and water services, as well as more private sector industries such as banking, airports, and transport. It is via these examples that we explained why academic methods should be preferred to numerous heuristic methods. Results from research for policy makers should be used only if the work is understandable and appropriate to the policy question being addressed, and can result in actionable decisions.

Apart from an efficient execution of public services, it is also important to assess whether the public sector does in fact provide the services that contribute to achieving substantial social goals.

The efficiency and effectiveness of public and social services is especially crucial given that the current austerity measures, the strains on all levels of government budgets, and the increasing importance of minimizing costs increase the demand for more research and the appropriate application of quantitative methods.

An effective and efficient execution of public services provides a number of social benefits. Utilizing a more rational approach starts from developing politically formulated goals. By using theories from economics, market organization, and regulation, effectiveness can be measured and reported on.

Throughout this book, we have provided not only the theoretical constructs of efficiency and productivity, but also simple examples and illustrations taken from the academic literature. It is our hope that this will encourage readers to pursue these types of analyses, both in the real world and in academia. We also hope that we have presented a compelling argument for the increased use of productivity analysis for the public and non-profit sector, and have encouraged analysts to develop the quantitative skills presented here. By presenting the results of thorough research, decision makers can be guided by findings that are objectively determined rather than on political biases.

To reiterate, this book is meant for everyone who is interested in a more effective and productive public sector, as well as for those who are interested in the performance of the public sector in general and public firms in particular. Policy-makers at departments and municipalities, politicians, representatives of umbrella organizations, unions, consumer organizations, members of boards, supervisors, managers of firms, and researchers and students in the field of public administration, public management, public policy analysis, and public finances can all glean something from this book. It is also meant to stimulate future researchers who wish to conduct the types of analyses presented here, especially as dynamic changes may be in the offing requiring solid research based on theoretical grounds may well be in the offing.

The complex statistics and mathematics that are common in productivity measurement were avoided wherever possible. Although the theory was explained based on simple examples and comprehensible figures, the reader was required to have some insight into basic mathematical equations and be able to grasp graphics.

Jos L.T. Blank and Vivian G. Valdmanis

Annex of Chapter 3

Table B3-1 Aspects representation production structure

Name	*Notation*	*First vector*	*Second vector*	*Economic behavior*	*Control variable*	*Constraint*
Direct input distance function	D_I (y,x)	ni, q-vex	nd, caaf, h	None	Resources	Production
Direct output distance function	D_O (x,y)	ni, q-caaf	nd, vex, h	None	Production	Resources
Indirect input distance function	ID_I (p/R,x)	nd, q-vex	nd, caaf, h	None	Resources + Production (partial)	Revenue
Indirect output distance function	ID_O (w/C,y)	nd, q-caaf	nd, vex, h	None	Production + Resources (partial)	Costs
Cost function	C (y,w)	nd, q-vex	nd, caaf, h	Cost minimization	Resources	Production
Revenue function	R (x,p)	nd, q-caaf	nd, vex, h	Revenue maximization	Production	Resources
Indirect revenue function	IR (w/C,p)	ni, q-vex	nd, vex, h	Revenue maximization	Production + Resources (partial)	Costs
Indirect cost function	IC (p/R,w)	ni, q-caaf	nd, caaf, h	Cost minimization	Resources + Production (partial)	Revenue
Profit function	Π (w,p)			Profit maximization	Resources+ Production	None

y = production; x = resources; p = product prices; w = resources prices; R revenue; C = costs.

nd = non-decreasing; ni = non-increasing; q-caaf = quasi-concave; q-vex = quasi-convex; caaf = concave; vex = convex; h = homogeneous of degree one.

Annex of Chapter 4

Definition of entropy index

The entropy index equals:

$$E = p_1 \ln(p_1) + p_2 \ln(p_2) + \cdots + p_M \ln(p_M)$$

(ln(.) is the natural logarithm)

And p_m is defined as:

$$p_m = \frac{y_M}{y} = \frac{y_M}{y_1 + y_2 + \cdots + y_M}$$

In which:
E = entropy;
p_m = service share belonging to group m (m = 1,... ,M);
y_m = number of services in group m.

Definition of Gini coefficient

The Gini coefficient equals:

$$G = \frac{2}{M^2 \overline{y}}(y_1 + 2y_2 + \ldots + My_M) - \frac{M+1}{M}$$

$$G = \frac{2}{M^2 \bar{y}}(y_1 + 2y_2 + \cdots + My_M) - \frac{M+1}{M}$$

With:

$\bar{y}$ = the number of services per group at an equal distribution over all groups.

For the example in paragraph 5.8.6 this means that the entropy index equals 0.56 (= 0.25 * ln (0.25) + 0.75 * ln (0.75)), the Gini coefficient equals 0.25 (=1 * (1 * 0.25 + 2 * 0.75) – 3/2) and the Herfindahl index equals 0.63 (= $0.25^2 + 0.75^2$).

Annex of Chapter 5

Transcendental logarithmic function
The transcendental logarithmic function (in short: translog) is (applied on a cost function) as follows:

$$\ln(C) = a_0 + \sum_{m=1}^{M} b_m \ln(y_m) + \sum_{n=1}^{N} c_n \ln(w_n) + \frac{1}{2}\sum_{m=1}^{M}\sum_{m'=1}^{M} b_{mm'} \ln(y_m) \ln(y_{m'}) + \frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N} c_{nn'} \ln(w_n) \ln(w_{n'}) + \sum_{m}^{M}\sum_{n}^{N} e_{mn}\ln(y_m) \ln(w_n)$$

In which:

C = costs;
y_m = product m (m = 1,... , M);
w_n = price of resource n (n = 1,... , N).

a_0, b_m, c_n, $b_{mm'}$ $c_{nn'}$ and e_{mn} are the, to be estimated, parameters of the function.

The translog function is a flexible one. This means that the function, depending on the parameter values, can take a wide variety of forms. For the parameter values, however, there are a number of conditions and constraints. The most important ones are:
Homogeneity of the degree 1 in prices:

$\sum_{n=1}^{N} c_n = 1$; $\sum_{n=1}^{N} c_{nn'} = 0$ $(\forall n')$; $\sum_{n}^{N} e_{mn} = 0$ $(\forall m)$
(Homogeneity of the degree 1 in prices)

$b_{mm'} = b_{m'm}$; $c_{nn'} = c_{n'n}$
(symmetry)

From the cost function we can deduct the so-called cost share functions.

$S_n = c_n + \sum_{n'=1}^{N} c_{nn'} \ln(w_{n'}) + \sum_{m}^{M} e_{mn} \ln(y_m)$ $(\forall n)$

From the translog we can also, quite simply, calculate values such as elasticity of scale, diversification effects, marginal costs and elasticity of substitution.

Notions and definitions

Allocative efficiency (input orientation)
Producing a certain level of services at minimal costs by using an optimal mix of resources at given resources prices and technology.

Allocative efficiency (output orientation)
Realizing maximum revenues by producing an optimal mix of services at given resources, product prices and technology.

Autonomous development/growth cost shares
The development of the average cost shares that remain after correction for the development of resource prices and service levels, mostly due to changes in technology, legislation and environment.

Autonomous productivity development/growth
The development of the services produced that remains after correction for the development of resource usage, mostly due to changes in technology, legislation and environment.

Budget restricted indirect output model
Economic model describing the relation between services produced, resource prices, budget and technical change.

Capital costs
Costs that relate to the utilization of capital. They usually are part of the fixed costs. Capital costs consist of the depreciation of capital and the opportunity cost of capital (reflected by interest rate).

Constant economies of scale
Situation in which a percentage increase of resources leads to an equiproportional percentage increase of services.

Contractual working time
The potential working time per year, reduced with the number of contractual days off, holidays and holiday shifts.

Cost flexibility
The percentage effect of a 1% increase of all services produced on the total costs.

Cost frontier
Set of firms that produce certain levels of services at lowest costs. Also indicated as best practice.

Cost function
A mathematical model describing the relation between the costs on the one hand, and produced service quantities and resource prices on the other.

Cost share
The portion of the costs of a certain resource in total costs.

Cost share equation
Equation that describes the optimal relation between a cost share of a certain resource on the one hand and produced services, the resource prices and technical changes on the other.

Data Envelopment Analysis
Analysis technique based on linear programming techniques to calculate the efficiency scores for each firm.

Depreciation
Estimated value of the part of the fixed assets that is used in the production process during a certain time period.

Diseconomies of scale
Situation in which a percentage growth of the resources leads to a less than proportionate percentage growth of services produced.

Economic efficiency
Technical and allocative efficiency combined.

Economies of scale
Situation in which a percentage growth of the resources leads to a more than proportionate percentage growth of services produced.

Economies of scope
Situation in which the joint production of services leads to cost savings in comparison to the separate production of these services.

Effectiveness
The extent to which (social) outcomes can be increased at given resources (or policy instruments) and available technology.

Efficiency
See allocative, economic, technical, total and scale efficiency.

Efficient by default
Being efficient because there are no firms with a similar or higher service level (at Data Envelopment Analysis). So no comparisons can be made.

Elasticity of demand
The percentage change in the use of a resource as a result of 1% change in relative resource prices.

Elasticity of scale
The percentage effect of a 1% increase of all resources on service levels.

Elasticity of substitution
The percentage change in the ratio of two resources as a result of a 1% change in the relative prices of these resources. Measures the responsiveness in the usage of resources of a firm to changes in relative resource prices.

Envelope condition
This condition determines the usage of capital at which long run and short run cost minimization coincides.

Fixed costs
Costs of the resources that cannot be changed on short term. Examples are depreciation and interest costs.

Fixed production
Part of the production that cannot be altered on the short run. An example is the polyclinic deliveries in a hospital, which cannot be influenced by the hospital management.

Fixed use of resources
Part of the resources, such as buildings, that is a short-term data in the production process.

Full time factor
Average number of employees per full time equivalent.

Herfindahl index (HHI)
Measure for the concentration of services at firms in a certain region. Number between 0 and 1. Often used as measurement for competition. A low number

corresponds to a very competitive market, a high number to monopolistic tendencies.

Input set
All combinations of resources with which given services can be achieved.

Insourcing
Executing activities for third parties (e.g. laundry services).

Isocost line
Combinations of quantities of resources that entail the same amount of costs.

Isoquant
The combinations of resources at which given services can be produced and no proportional reduction of all resources is possible.

Labor costs
Costs that relate to the employment of labour. These costs are mostly indicated as variable costs.

Labor productivity
Ratio between service volume and staff volume.

Marginal costs
The costs that result from the production of one extra unit of a certain product type.

Material supplies
In most research material supplies is used for everything that cannot be indicated as staffing costs or capital costs. The employment of material comprises, among others: allocation maintenance, minor repairs and exploitation, energy and water, housekeeping, levies, inventory, equipment and training aids and other non-staffing and non-capital costs.

Material costs
Costs that relate to the utilization of material supplies.

Occupation rate
Indicates the degree in which a resource is used. Is mostly related to capital goods. Example is the occupation rate of hospital beds: the average number of nurses per acknowledged bed, expressed in the total amount of days in a year.

Optimal cost share
The cost shares of resources that lead to maximum output. These can be calculated with the budget indirect output model.

Outcome
Measures for the extent to which social or public goals are achieved.

Output
See production.

Production volume or volume of the production
Standard for the volume of the supplied end products. When there is more than one service this is a weighted sum of the distinct services.

Outsourcing
Having activities carried out by third parties (e.g. laundry services).

Overhead
Many definitions are possible. Most commonly used is the employment of resources that are not directly linked to the primary process. Examples of overhead are management, supporting staff, and material supplies.

Overutilization
When the cost share of a certain resource of a school is higher than the optimal cost share of this resource.

Panel data
Data of individual measurement units (firms or people) over multiple time periods.

Partial productivity
Ratio between the service volume and one of the resources. Example is labor productivity.

Production
Level of the produced products and services.

Production function
A mathematical model describing the (technical) relation between services and resources.

Productivity
See total productivity.

Regional price
The average price of a resource in a region.

Returns to scale
The percentage change of the services that occurs as a result of a percentage change of all resources. Mostly expressed as a number around 1. If the number is smaller than 1 then there are decreasing returns to scale, if greater than 1 then there are increasing returns to scale.

Scale effect
See also returns to scale. Positive scale effects are associated with increasing returns to scale, negative scale effects with decreasing returns to scale.

Scale efficiency
Measure for the percentage loss of production due to not producing at the optimal scale.

Shadow costs
The costs calculated based on the shadow prices.

Shadow prices
The (virtual) prices at which the observed mix of resources is optimal assuming the firm is cost minimizing..

Social costs
Total of costs that relate to the production of services and goods. These include direct business costs or internal costs, but also costs related to the acquirement of the service. In education, for instance, it is not only about the costs for school attendance, but also about the costs for transport of the students to and from school and the efforts of parents to help their children with their homework..

Staff volume
Measure for the total use of different types of employees.

Staffing costs
Costs related to the use of labor (so, including social security and such).

Synergy effect
The costs savings that occur at a combined supply of various product types by one producer, in relation to the situation in which a similar production is

achieved by individual producers, who each specialize in supplying one product type.

Technical change
Relates to productivity changes due to the overall process of invention, innovation and diffusion of technology.

Technical efficiency
Measure that reflects the extent to which services can be (radially) expanded without adding extra resources.

Total costs
Sum of all costs related to the usage of resources in the production process. This includes labor costs, material costs (including paid indirect taxes) and capital costs (depreciation and interest payments).

Total factor productivity
Ratio between the service volume and the volume of resources.

Underuitilization
When the cost share of a certain resource is lower than the optimal cost share of this resource.

Variable costs
The costs of the resources that can be adjusted to changing circumstances by the producer on short term. This contrary to the fixed costs, such as depreciation and interest payments that are (practically) fixed for a number of years.

Volume of the resources
Measure for the use of all different type of resources.

Working year
The use of one full time employee for the duration of one year, or a similar use of part time employees. Two employees who work half time count, in principle, as one working year.

List of tables

List of figures

References

Afonso, A., & St. Aubyn, M. (2006). Cross-country efficiency of secondary education provision: A semi-parametric analysis with non-discretionary inputs. *Economic Modelling*, *23*, 476–491.

Aigner, D., Lovell, C. A. K., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, *6*(1), 21–37. http://doi.org/10.1016/0304-4076(77)90052-5

Alam, I. S., & Granderson, G. (2008). Organizational structure and productive efficiency of non-profit hospitals. In J. L. T. Blank & V. G. Valdmanis (Eds.), *Evaluating hospital policy and performance: contributions from hospital policy and productivity research* (Vol. 18, pp. 63–88). Oxford: Elsevier.

Alvarez, A., Amsler, C., Orea, L., & Schmidt, P. (2006). Interpreting and testing the scaling property in models where inefficiency depends on firm characteristics. *Journal of Productivity Analysis*, *25*(3), 201–212.

Balk, B. M. (2003). The Residual: On Monitoring and Benchmarking Firms, Industries, and Economies with Respect to Productivity. *Journal of Productivity Analysis*, *20*, 5–47.

Banker, R. D., A. Charnes, & Cooper, W. W. (1984). Some models for estimating technical and scale efficiencies in Data Envelopment Analysis. *Management Science*, *30*(9), 1078–1092.

Barnum, D. T., McNeil, S., & Hart, J. (2007). Comparing the efficiency of public transportation subunits using Data Envelopment Analysis. *Journal Of Public Transportation*, *10*(2), 1–16.

Bates, L. J., Mukherjee, K., & Santerre, R. E. (2006). Market Structure and Technical Efficiency in the Hospital Services Industry: A DEA approach. *Med Care Res Rev*, *63*(4), 499–524. http://doi.org/10.1177/1077558706288842

Battese, G. E., & Coelli, T. J. (1995). A model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data. *Empirical Economics*, *20*(2), 325–332.

Baumol, J., Panzar, J. C., & Willig, R. D. (1988). *Contestable markets and the theory of industry structure* . Sydney: Marcourt Brace Jovanovich.

Bazzoli, G. J. (2008). Hospital consolidation and integration activity in the United States. In J. L. T. Blank & V. G. Valdmanis (Eds.), *Evaluating hospital policy and performance: contributions from hospital policy and productivity research* (Vol. 18, pp. 45–61). Oxford: Elsevier.

Benbassat, J., & Taragin, M. (2000). Hospital readmissions as a measure of quality of health care: advantages and limitations. *Archives of Internal Medicine*. Retrieved from

http://jamanetwork.com/journals/jamainternalmedicine/fullarticle/415392
Blank, J. L. T. (1993). *Kosten van kennis: een empirisch onderzoek naar de productiestructuur van het basisonderwijs in Nederland*. *Sociale en Culturele Studie*. Rijswijk/Den Haag: SCP/VUGA.
Blank, J. L. T. (2000). *Public provision and performance: contributions from efficiency and productivity measurement*. Amsterdam: Elsevier.
Blank, J. L. T. (2008). Innovations and Productivity: An empirical investigation in Dutch Hospital Industry. In J. L. T. Blank & V. G. Valdmanis (Eds.), *Evaluating hospital policy and performance: contributions from hospital policy and productivity research* (Vol. 18, pp. 89–109). Oxford: Elsevier.
Blank, J. L. T. (2009). Non-maximizing output behavior for firms with a cost-constrained technology. *Journal of Productivity Analysis*, *31*(1), 27–32.
Blank, J. L. T. (2013). Maximizing public value for subsidized non-profit firms: a mathematical economic model. *Journal of Productivity Analysis*, *40*(173), 1–11. http://doi.org/10.1007/s11123-012-0318-7
Blank, J. L. T. (2015). Benchmarking and operational management: an application of frontier analysis to Dutch secondary education. In K. De Witte (Ed.), *Contemporary Economic Perspectives in Education* (pp. 77–88). Leuven: Leuven University Press.
Blank, J. L. T., & Lovell, C. (2000). Performance assessment in the public sector: contributions from efficiency and productivity measurement. In J. L. T. Blank (Ed.), *Public provision and performance*. Amsterdam: Elsevier.
Blank, J. L. T., & Valdmanis, V. G. (2008). *Evaluating hospital policy and performance: contributions from hospital policy and productivity research*. Oxford: Elsevier JAI. http://doi.org/BO0701
Blank, J. L. T., & van Hulst, B. L. (2005). Verspreiding van vernieuwing. In RVZ (Ed.), *Weten wat we doen - Verspreiding van innovaties in de zorg*. Zoetermeer: Raad voor de Volksgezondheid en Zorg.
Blank, J. L. T., & van Hulst, B. L. (2009). Productive Innovations in Hospitals: An Empirical Research on the Relation between Technology and Productivity in the Dutch Hospital Industry. *Health Economics*, *18*(3), 665–679.
Blank, J. L. T., van Hulst, B. L., Koot, P. M., & van der Aa, R. (2012). Benchmarking overhead in education: a theoretical and empirical approach. *Benchmarking: An International Journal*, *19*(2), 239–254. http://doi.org/http://dx.doi.org/10.1108/14635771211224554
Bruijn, H. de. (2007). *Managing performance in the public sector*. London: Routledge.
Button, K., & Costa, A. (1999). Economic efficiency gains from urban public transport regulatory reform: Two case studies of changes in Europe. *The Annals of Regional Science*, *33*(1999), 425–438.
Caves, D. W., Christensen, L. R., & Diewert, W. E. (1982). The economic-theory of index numbers and the measurement of input, output, and

productivity. *Econometrica*, *50*(6), 1393–1414. http://doi.org/10.2307/1913388

Charnes, A., Cooper, W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*(6), 429–444. http://doi.org/10.1016/0377-2217(78)90138-8

Chilingerian, J. A. (1995). Evaluating physician efficiency in hospitals: a multivariate-analysis of best practices. *European Journal of Operational Research*, *80*(3), 548–574. http://doi.org/10.1016/0377-2217(94)00137-2

Cobb, C. W., & Douglas, P. H. (1928). A theory of production. *American Economic Review*, *18 (Supple*, 139–165.

Coelli, T. J., & Fleming, E. (2004). Diversification economies and specialisation efficiencies in a mixed food and coffee smallholder farming system in Papua New Guinea. *Agricultural Economics*, *31*(2–3), 229–239. Retrieved from http://www.sciencedirect.com/science/article/B6T3V-4DN1HVS-1/2/d271ced508efa689dedac8d80eb8870b

Coelli, T. J., Rao, D. S. P., O'Donnell, C. J., & Battese, G. E. (2005). *An Introduction to efficiency and productivity analysis* (2nd ed.). New York: Springer.

Cooper, W. W., L.M. Seiford, & K. Tone. (2007). *Introduction to Data Envelopment Analysis and Its Uses: With DEA Solver Software and References*. New York: Springer.

Cullmann, A., Crespo, H., & Plagnet, M. A. (2008). *International Benchmarking in Electricity Distribution: A comparison of French and German utilities*. Berlin: Dept. of International Economics.

Davies, S., Saynina, O., & Schultz, E. (2013). Implications of metric choice for common applications of readmission metrics. *Health Services*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/1475-6773.12075/full

Debreu, G. (1951). The Coefficient of Resource Utilization. *Econometrica*, *19*(3), 273–292. Retrieved from http://www.jstor.org/stable/1906814

Deloitte Research. (2004). *Mastering innovation: Exploiting ideas for profitable growth*. (D. Consulting, Ed.). New York: Deloitte Consulting.

Dervaux, B., Ferrier, G. D., Leleu, H., & Valdmanis, V. G. (2004). Comparing French and US hospital technologies: A directional input distance function approach. *Applied Economics*, *36*(10), 1065–1081.

Dervaux, B., Leleu, H., Minvielle, E., Valdmanis, V. G., Aegerter, P., & Guidet, B. (2009). Performance of French intensive care units: A directional distance function approach at the patient level. *International Journal of Production Economics*, *120*(2), 585–594. http://doi.org/10.1016/j.ijpe.2009.04.011

Dervaux, B., Leleu, H., Nogues, H., & Valdmanis, V. G. (2006). Assessing French nursing home efficiency: An indirect approach via budget-constrained DEA models. *Socio-Economic Planning Sciences*, *40*(1), 70–91. http://doi.org/DOI: 10.1016/j.seps.2004.10.005

Dijkgraaf, E., & Varkevisser, M. (2007). *Analyse kosteneffect fusies drinkwatersector*. Rotterdam: Erasmus Universiteit Rotterdam.

Fan, Y., Li, Q., & Weersink, A. (1996). Semiparametric estimation of stochastic production frontier models. *Journal of Business & Economic Statistics*, *14*(4), 460–468.

Färe, R., Grosskopf, S., & Lovell, C. A. K. (1986). Scale economies and duality. *Journal of Economics-Zeitschrift Fur Nationalokonomie*, *46*(2), 175–182. http://doi.org/10.1007/bf01229228

Färe, R., Grosskopf, S., & Lovell, C. A. K. (1994). *Production Frontiers*. Cambridge: Cambridge University Press.

Färe, R., Grosskopf, S., Norris, M., & Zhang, Z. (1994). Productivity Growth, Technical Progress, and Efficiency Change in Industrialized Countries. *The American Economic Review*, *84*(1), 66–83. http://doi.org/10.2307/2117971

Färe, R., & Primont, D. (1995). *Multi-Output Production and Duality: Theory and applications*. Dordrecht: Kluwer Academic Publishers.

Farrell, M. J. (1957). The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society. Series A (General)*, *120*(3), 253–290. Retrieved from http://www.jstor.org/stable/2343100

Fitzpatrick, T., & McQuinn, K. (2008). Measuring Bank Profit Efficiency. *Applied Financial Economics*, *18*(1–3), 1–8.

Fizel, J. L., & Nunnikhoven, T. S. (1993). The Efficiency of Nursing Home Chains. *Applied Economics*, *25*(1), 49–55. http://doi.org/10.1080/00036849300000112

Fried, H. O., Lovell, C. A. K., & Schmidt, S. S. (2008). *The measurement of productive efficiency and productivity growth*. New York: Oxford University Press.

Goes, J., & Zhan, C. (1995). The effects of hospital-physician integration strategies on hospital financial performance. *Health Services Research*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1070073/

Greene, W. (2004). Distinguishing between heterogeneity and inefficiency: Stochastic Frontier Analysis of the World Health organization's panel data on national health care systems. *Health Economics*, *13*, 959–980.

Grosskopf, S., Hayes, K. J., Taylor, L. L., & Weber, W. L. (2000). Input regulations and allocative efficiency in U.S. public schools. In J. L. T. Blank (Ed.), *Public Provision and Performance* (pp. 175–190). Amsterdam: Elsevier.

Haelermans, C., & Blank, J. (2012). Is a schools performance related to technical change? a study on the relationship between innovations and secondary school productivity. *Computers & Education*, *59*, 884–892.

Haelermans, C., De Witte, K., & Blank, J. (2012). On the allocation of resources for secondary schools. *Economics of Education Review*, *31*(5), 575–586. http://doi.org/10.1016/j.econedurev.2012.02.007

Herrin, J., Andre, J. S., & Kenward, K. (2015). Community factors and hospital readmission rates. *Health Services*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/1475-6773.12177/full

Hulten, C., Dean, E., & Harper, M. (2001). *New Developments in Productivity Analysis*. (NBER, Ed.)*Studies in Income and Wealth* (Vol. 63). Chicago: The University of Chicago Press.

Johnson, A., & Kuosmanen, T. (2015). An Introduction to CNLS and StoNED Methods for Efficiency Analysis: Economic Insights and Computational Aspects. *Benchmarking for Performance Evaluation*. Retrieved from http://link.springer.com/chapter/10.1007/978-81-322-2253-8_3

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. http://doi.org/10.1016/0304-4076(82)90004-5

Kaplan, R. S., & Norton, D. P. (1996). *The balanced scorecard: Translating strategy into action*. Boston: Harvard Business School Press.

Khan, S. U. (2005). *Macro Detereminants of Total Factor Productivity in Pakistan*. *Munich Personal RePEc Archive* . Munich: University of Munich. Retrieved from http://mpra.ub.uni-meunchen.de/8693/

Kocher, R. P., & Adashi, E. Y. (2011). Hospital readmissions and the Affordable Care Act: paying for coordinated quality care. *Jama*, *306*(16), 1794–1795.

Koopmans, T. C. (1951). An analysis of production as efficient combination of activities. In T. C. Koopmans (Ed.), *Cowless commission for research in economics, monograph no. 13*. New York: John Wiley and Sons, Inc.

Kooreman, P. (1994). Nursing home care in the Netherlands: A nonparametric efficiency analysis. *Journal of Health Economics*, *13*(3), 301–316. http://doi.org/doi:10.1016/0167-6296(94)90029-9

Kumbhakar, S. C., & Lovell, C. (2000). *Stochastic frontier analysis*. New York: Cambridge University Press.

Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. *The Econometrics Journal*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1368-423X.2008.00239.x/full

Lensink, R., Meesters, A., & Naaborg, I. (2008). Bank Efficiency and Foreign Ownership: Do Good Institutions Matter? *Journal of Banking and Finance*, *32*(5), 834–844.

Lovell, C. A. K. (2000). Measuring efficiency in the public sector. In J. L. T. Blank (Ed.), *Public provision and performance: contributions from efficiency and productivity measurement* . Amsterdam: Elsevier.

Martın, J. C., Roma, C., & Voltes-Dorta, A. (2009). A Stochastic Frontier Analysis to estimate the relative efficiency of Spanish airports. *Journal of Productivity Analysis*, *31*(3), 163–176.

McLaughlin, K., Osborne, S. P., & Ferlie, E. (2002). *New Public Management: Current Trends and Future Prospects* . London: Routledge.

Mutter, R. L., & Rosko, M. D. (2008). The impact of ownership on the cost-efficiency of U.S. hospitals. In J. L. T. Blank & V. G. Valdmanis (Eds.), *Evaluating hospital policy and performance: contributions from hospital policy and productivity research* (Vol. 18, pp. 113–138). Oxford: Elsevier.

Niaounakis, T. K., & Blank, J. L. T. (2015). *Lasten van (samen) belasten. Een empirisch onderzoek naar de doelmatigheid van de gemeentelijke belastingheffing en de uitvoering van de Wet WOZ tussen 2005 en 2012* (IPSE Studies Research Reeks). Den Haag/Delft: IPSE Studies.

Niaounakis, T. K., & Blank, J. L. T. (2017). Inter-municipal cooperation, economies of scale and cost efficiency: an application of stochastic frontier analysis to Dutch municipal tax departments. *Local Government Studies*. Retrieved from http://www.tandfonline.com/doi/full/10.1080/03003930.2017.1322958

Niskanen, W. A. (1968). The Peculiar Economics of Bureaucracy. *American Economic Review*, *57*(2), 293–321.

Nuckols, T. K. (2015). County-Level Variation in Readmission Rates: Implications for the Hospital Readmission Reduction Program's Potential to Succeed. *Health Services Research*, *50*(1), 12–19. http://doi.org/10.1111/1475-6773.12268

OECD. (2012a). % total expenditure on health by provider. Paris: OECD. Retrieved from http://stats.oecd.org/BrandedView.aspx?oecd_bv_id=health-data-en&doi=health-data-en

OECD. (2012b). OECD Health Data 2011. OECD.

Onderwijsraad. (2008). *De bestuurlijke ontwikkeling van het Nederlandse onderwijs*. Den Haag: Onderwijsraad.

Oude Lansink, A., Reinhard, S., Jong, A. de, Luijt, J., & Wijk, M. van. (2008). *Efficiëntie van bosbedrijven: Data Envelopment Analysis (DEA)*. Den Haag: LEI.

Ozcan, Y. A. (2008). *Health care benchmarking and performance evaluation. An assessment using data envelopment analysis (DEA)*. International Series in Operations Research and Management Science. New York: Springer. Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=econ&AN=0979016

Parmeter, C., & Kumbhakar, S. (2014). *Efficiency Analysis: A Primer on Recent Advances*. Miami, New York.

Plagnet, M. A. (2006). *Use of benchmarking methods in Europe in the electricity distribution sector*. Electricité de France (EDF).

Ray, S. C., Kumbhakar, S. C., & Dua, P. (2015). *Benchmarking for performance evaluation: A production frontier approach*. *Benchmarking for Performance Evaluation: A Production Frontier Approach*. http://doi.org/10.1007/978-81-322-2253-8

Rodríguez-Álvarez, A., & Lovell, C. A. K. (2004). Excess capacity and expense preference behaviour in National Health Systems: an application to the Spanish public hospitals. *Health Economics*, *13*, 157–169.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New Jersey: John Wiley and Sons.

Sari, N. (2008). Competition and market concentration. In J. L. T. Blank & V. G. Valdmanis (Eds.), *Evaluating hospital policy and performance: contributions from hospital policy and productivity research* (Vol. 18, pp. 139–156). Oxford: Elsevier.

Schmidt, P. (2011). One-step and two-step estimation in SFA models. *Journal of Productivity Analysis*, *36*(2), 201–203. http://doi.org/10.1007/s11123-011-0228-0

Shephard, R. W. (1953). *Theory of Cost and Production Functions*. (P. U. Press, Ed.). Princeton: Princeton University Press.

Sherman, H. D., & Zhu, J. (2006). *Service Productivity Management: Improving Service Performance Using Data Envelopment Analysis (DEA)*. Includes DEAFrontier Software. New York: Springer. Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=econ&AN=0889140

Simar, L., Lovell, C. A. K., & Vanden Eeckaut, P. (1994). *Stochastic frontiers incorporating exogenous influences on efficiency* (Discussion Papers No. 9403). Institut de Statistique, Université Catholique de Louvain.

Solow, R. M. (1987). We'd better watch out. *New York Times Book Review*, (12), 36.

Tjosvold, D. (1988). Cooperative and competitive interdependence - collaboration between departments to serve customers. *Group & Organization Studies*, *13*(3), 274–289. http://doi.org/10.1177/105960118801300303

Triebs, T. P., Saal, D. S., Arocena, P., & Kumbhakar, S. C. (2012). *Estimating Economies of Scale and Scope with Flexible Technology*. *Ifo Working Paper*. Munich: Leibniz Institute for Economic Research. Retrieved from http://www.cesifo-group.de/portal/pls/portal/docs/1/1219550.PDF

Tulder van, F. (1994). *Van misdaad tot straf: een economische benadering van de strafrechtelijke keten*. Den Haag: VUGA/Sociaal en Cultureel Planbureau.

Valdmanis, V., DeNicola, A., & Bernet, P. (2015). Public health capacity in the provision of health care services. *Health Care Management Science*, *18*(4), 475–482. http://doi.org/10.1007/s10729-014-9277-z

Valdmanis, V. G., Bernet, P., & Moises, J. (2010). Hospital Capacity, Capability, and Emergency Preparedness. *European Journal of Operations Research*, *207*(3), 1628–1634.

Valdmanis, V. G., Rosko, M. D., & Mutter, R. L. (2008). Hospital quality, efficiency, and input slack differentials. *Health Services Research*, *43*(5), 1830–1848. http://doi.org/10.1111/j.1475-6773.2008.00893.x van Hulst, B. L. (2000). *De incidentele loonontwikkeling in de algemene en categorale ziekenhuizen, 1994-1999*. Utrecht: Prismant.

Vandermeulen, L. J. R., van Hulst, B. L., & Mollerus, A. C. (1997). *De incidentele loonontwikkeling in de NZf-sectoren*. Utrecht: Nederlandse Zorgfederatie.

Visser, J., Arnold, E., & Pieter, D. (2005). *De incidentele loonontwikkeling in de algemene en categorale ziekenhuizen, 2000-2003* . Utrecht: Prismant.

Wang, H. J., & Schmidt, P. (2002). One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. *Journal of Productivity Analysis*, *18*(2), 129–144. http://doi.org/10.1023/A:1016565719882

Weiss, M., Yakusheva, O., & Bobay, K. (2011). Quality and cost analysis of nurse staffing, discharge preparation, and postdischarge utilization. *Health Services Research*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1475-6773.2011.01267.x/full