Charles University in Prague

Faculty of Social Sciences
Institute of Economic Studies

MASTER'S THESIS

The Efficiency of Public Expenditure, Evidence from the Czech Republic

Author: **Bc. Veronika Vraná**

Supervisor: **Jan Zápal, Ph.D.**

Academic Year: **2015/2016**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain a different or the same degree.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, May 9, 2016

Signature

Acknowledgments

The author is grateful especially to Jan Zápal, Ph.D. for the extensive support and valuable advice during the whole process of writing, to PhDr. Eva Vacková for her comments on grammar and stylistics, and to numerous civil servants for their help with the study of regional services and search for corresponding data.

Abstract

Efficiency is an important concept for performance evaluation of decision-making units. This thesis studies efficiency of public spending and methods of its estimation. Firstly, a wide range of efficiency estimators are defined and compared. For the public spending efficiency analysis the most convenient estimator is found to be the cost frontier approach of the Stochastic Frontier Analysis (SFA).

The second part of the thesis contains efficiency analysis of public spending in the 14 regions of the Czech Republic in the years 2003–2014. For the analysis current expenditure of regional offices is used as the input. Based on the investigation of regional services output index is formed and employed in the analysis. The estimation is performed using the cost frontier approach of the SFA as the main method. Various other efficiency estimators are then applied to the data in order to study robustness of the results. The thesis further discusses relative ranking of the regions and time evolution of efficiency scores using different assumptions and variety of methods. Lastly, several potential effects on the estimated efficiency are considered and analysed.

JEL Classification	D24, H41, H72
Keywords	efficiency, regions, public expenditure, Czech Republic

Author's e-mail	`verca25.vrana@seznam.cz`
Supervisor's e-mail	`j.zapal@cerge-ei.cz`

Abstrakt

Práce se zabývá efektivitou veřejných výdajů a metodami odhadů efektivity. Metody jsou v práci popsány a porovnány podle vhodnosti pro analýzu efektivity obecně a konkrétně pro analýzu efektivity veřejných výdajů. Jako nejvhodnější metoda je vybrána stochastická obálková analýza se zaměřením na odhad výdajové hraniční funkce. Tuto metodu, spolu s dalšími alternativními metodami odhadů pro porovnání výsledků, používá následná analýza efektivity veřejných výdajů v krajích České republiky. Analýza využívá panelových dat pro roky 2003–2014. Jako vstup jsou na základě dřívějších prací vybrány běžné výdaje krajského úřadu. Jako výstup je pro účely této práce vytvořen index zachycující poskytování krajských veřejných služeb, který je zkonstruován na základě zkoumání krajských služeb. Pořadí krajů podle různých metod a předpokladů je okomentováno a porovnáváno. Poslední část práce zkoumá vlivy na efektivitu.

Klasifikace JEL	D24, H41, H72
Klíčová slova	efektivita, kraje, veřejné výdaje, Česká republika

E-mail autora	verca25.vrana@seznam.cz
E-mail vedoucího práce	j.zapal@cerge-ei.cz

Contents

List of Tables

List of Figures

Acronyms

CNLS Convex Nonparametric Least Squares

COLS corrected ordinary least squares

CRS constant returns to scale

DEA Data Envelopment Analysis

DFA deterministic frontier analysis

DMU decision-making unit

FDH Free Disposal Hull

FE Fixed effects

KH95 Kumbhakar & Heshmati (1995)

KLH14 Kumbhakar *et al.* (2014)

MLE maximum likelihood estimator

OLS ordinary least squares

NIRS non-increasing returns to scale

PFA Partial Frontier Analysis

PPF productivity possibility frontier

RE Random effects

SFA Stochastic Frontier Analysis

StoNED Stochastic Non-parametric Envelopment of Data

VRS variable returns to scale

Master's Thesis Proposal

Author	Bc. Veronika Vraná
Supervisor	Jan Zápal, Ph.D.
Proposed topic	The Efficiency of Public Expenditure, Evidence from the Czech Republic

Motivation Efficiency of public spending is one of the key problems in each country or region. If resources are spent efficiently more can be accomplished or spare resources may serve a different purpose. However, to approach efficiency it has to be measured whether and where the inefficiency occurs. The theoretical literature offers several methods to analyse efficiency. Each of them displays some advantages and disadvantages.

This thesis builds on Afonso, Schuknecht & Tanzi (2005 & 2010) and their application of three main methods of efficiency measurement (composite indicators, Data Envelopment Analysis, Free Disposal Hull) to a set of European countries. Along with the methods used by Afonso et al. other approaches are compared, such as in Coelli, Rao, Battese (2005). The inspiration to focus on regions instead of countries came from De Borger & Kerstens (1994) who used Belgian data on municipalities to assess the efficiency even on lower statistical units.

Hypotheses

Hypothesis #1: There are significant differences between different regions and Prague is the most efficient region.

Hypothesis #2: There are significant differences between the different methods of efficiency analysis where the Free Disposal Hull recognizes most of the regions as efficient while Data Envelopment Analysis finds only few to be efficient.

Hypothesis #3: After entering the European Union the efficiency decreased.

Methodology The first part of the thesis will describe different methods used to assess the efficiency of public spending. These include for example Data Envelopment Analysis method, as well as Free Disposal Hull or Stochastic Frontier methods. These methods will be first described and compared theoretically and then used on Czech data. The paper seeks to analyse the efficiency of public spending in the 14 Czech administrative regions. The goal is to compare the results not only among the regions, but also across the different methods. If data allows, the results could be used to compare efficiency before and after the Czech Republic entered the European Union.

Expected Contribution The thesis plans to summarize and describe methods of efficiency analysis that are currently used in the literature. It seeks to compare these methods theoretically as well as practically – in the analysis of spending efficiency in Czech regions. There have not yet been many analyses treating the efficiency of public spending in the Czech Republic and most of these focus just on education sector while this thesis aims to analyse the overall efficiency. While specialization offers a deeper insight into what measures should be taken to improve the efficiency of spending in particular regions, the general approach allows the comparison across the Czech Republic on a higher level. It is not common as well to choose regions as the units of interest as most papers focus on the whole countries thus not taking into account possible large difference between regions.

It is expected that this work will contribute to the theoretical literature on the methods of efficiency assessment for smaller geographical units than countries, as well as to the research by comparing Czech regions historically and geographically. The last part of the analysis – the possible change of efficiency due to accession to European Union – is meant to inform whether the potential inefficiency caused by access to additional resources is offset by higher fiscal control.

Outline

1. Introduction
2. Theoretical background
3. Definition of efficiency
4. Different ways to measure public spending efficiency
5. Comparison of methods
6. Analysis in the Czech Republic
7. Description of situation (public spending - composition, reforms)
8. Description of data

9. Analysis
10. Results
11. Application
12. Conclusion

Core bibliography

Afonso, A., Schuknecht, L., & Tanzi, V., (2005). "Public Sector Efficiency: An international comparison", Public Choice, Springer, vol. 123(3), pages 321–347, June.

Afonso, A., Schuknecht, L., & Tanzi, V. (2010) "Public sector efficiency: evidence for new EU member states and emerging markets", Applied Economics, vol. 42(17), pages 2147–2164

Afonso, A. and Fernandes, S. (2006) "Local government spending efficiency: DEA evidence for the Lisbon region", Regional Studies, 40, 39–53.

De Borger, B., & Kerstens, K. (1994). "Cost efficiency of Belgian local governments: A comparative analysis of FDH, DEA and econometric approaches," SESO Working Papers 1994002, University of Antwerp, Faculty of Applied Economics.

Coelli, Rao, O'Donnell & Battese (2005) "An Introduction to Efficiency and Productivity Analysis", 2nd edition, Kluwer Academic Publishers, Massachusetts.

Mandl, U., Dierx, A., & Ilzkovitz, F. (2008). "The effectiveness and efficiency of public spending" (No. 301). Directorate General Economic and Monetary Affairs (DG ECFIN), European Commission.

Sutherland, D., Price, R., Joumard, I., & Nicq, C. (2007). "Performance Indicators for Public Spending Efficiency in Primary and Secondary Education". OECD. Economics Department Working Paper, No. 546

Author Supervisor

Chapter 1

Introduction

Efficiency is a concept constructed to assess the performance of decision-making units and to detect their suboptimal use of resources. Due to scarcity of resources, efficiency is an important goal not only for private firms but also for the public sector, which needs to have enough means for the provision of services to citizens. When efficiency is increased, public goods and services can be provided in wider range and better quality.

This thesis concentrates on public spending efficiency and studies the efficiency analysis as a tool for ranking as well as for investigation of the reasons behind inefficiency. Efficiency analysis is based on the concept of Pareto-Koopmans efficiency. The original definition compares the units with their theoretically optimal counterpart. In practise, units are compared to each other. Thus a unit is efficient when no other unit produces more of all outputs without using more of at least one input and no other unit uses less of all inputs without producing less of at least one output.

An efficiency analysis of public sector brings various problematic aspects. It is more complicated to formulate objectives of decision-making units from the public sector. Appropriate data on input and output use are less available. It may be even arguable to what extent public organizations search for an optimal solution and what are their criteria of optimality. Yet similar to private firms, organizations from public sector have constrained budget. Additionally, most of their services are mandatory, at least to some extent. Thus, even though they do not face the same competition as private firms, public organizations still need not use their resources excessively and so the efficiency analysis of a regional public sector management seems to be useful.

Efficiency may be estimated using various techniques. Two most important groups in the current analysis are formed by deterministic non-parametric methods and stochastic parametric methods. The former techniques, which include Data Envelopment Analysis, Free Disposal Hull, and Partial Frontier Analysis, originated with the current concept of efficiency measurement. Thus the techniques are simpler, allowing the use of large numbers of inputs and outputs while imposing only few assumptions. Later adjustments adapt the estimators for better application by relaxing assumptions and by incorporating the estimators in statistical framework to allow statistical inference. The second group of estimators, which contains mainly Stochastic Frontier Analysis, allows naturally for inference since it is built on econometric estimation. By imposing various models, these techniques provide a wide range of estimators from more parsimonious with easier estimation to highly flexible ones that separate permanent inefficiency from its time-variation.

The discussion of the efficiency estimators in this thesis includes definitions of the methods, some of their extensions, and their comparison for general application and, in particular, for the public spending efficiency analysis. While the two groups of estimators may be generally viewed as complements, for the current analysis the cost frontier approach of Stochastic Frontier Analysis is chosen as the most appropriate. Robustness of the results is then studied using various alternative techniques from both groups of estimators.

This analysis uses data on all 14 Czech regions, including the capital city Prague. To increase the number of observations and enable the study of time-variation, data from years 2003–2014 are considered. As input variable current expenditure of regional offices is used. The thesis attempts to study overall efficiency, therefore the output side reflects the set of public services provided by regions. In order to avoid loss of power to reject efficiency, indicators of regional services are combined in the form of an output index.

Apart from ranking of regions and discussion of the relative efficiency the aim of this analysis is to study the causes of inefficiency. Several exogenous conditions are investigated for their potential effect on efficiency. Efficiency of Czech regions appears to be affected positively by GDP per capita, by higher number of parties in the regional council, and by higher share of communists and social democrats in the regional council. Higher density of population and higher average age seem to have a negative effect and no effect is found for the percentage of low educated population.

The thesis is structured as follows: the second chapter introduces the field of efficiency analysis by explaining definitions and structuring the estimators.

Chapter 3 covers the non-parametric deterministic methods for efficiency estimation. The parametric methods, along with a recent non-parametric stochastic method, are discussed in chapter 4. The theoretical part is concluded by a comparison of the methods at the end of chapter 4. Chapter 5 describes the data and justifies their selection. Chapter 6 contains results of the efficiency analysis of regions. The results are presented in the form of efficiency scores and a discussion of potential effects on efficiency. Chapter 7 concludes the thesis.

Chapter 2

Theoretical Background

This chapter presents the theoretical setting of the efficiency analysis. It begins with a definition of efficiency[1] and different efficiency perspectives: input or output efficiency and technical or allocative efficiency. It briefly discusses the difference between effectiveness and efficiency and mentions specific issues of public sector analyses at the end of the first section. The second section introduces the methods to estimate efficiency and their classification. The methods are defined in the following chapters.

2.1 Concept of efficiency

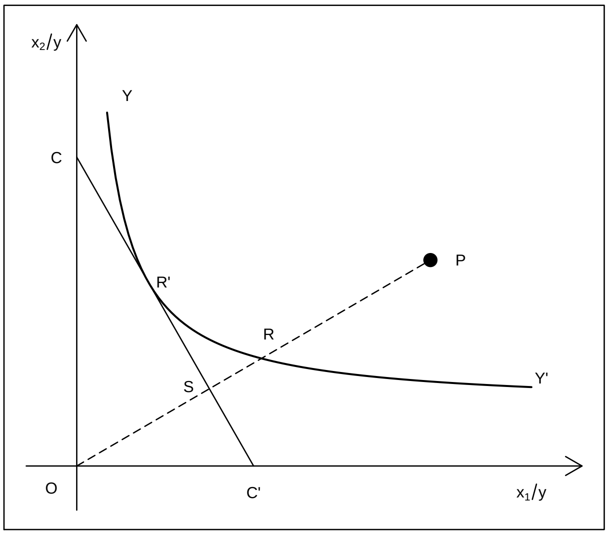
The efficiency and its measurement have been studied by a large body of literature since Farrell (1957) proposed and applied a way to measure not only the performance but also the efficiency of decision-making units (DMU). Some literature uses the expressions "performance" and "efficiency" interchangeably, but in most studies we find a clear distinction. While performance is simply the ratio of outputs to inputs, efficiency is understood to be the situation when resources are used in the production optimally. The optimality can be accomplished using two targets – maximising the output or minimising the input - which distinguishes two notions of efficiency, the input and output efficiency. The output efficiency describes a state when from a given amount of inputs the maximal possible amount of outputs is produced. On the other hand, the

[1] In this thesis the term "efficiency" is used to denote the Pareto-Koopmans efficiency which defines the efficiency so that "[a] DMU is fully efficient if and only if it is not possible to improve any input or output without worsening some other input or output (Cooper *et al.* 2007, p.45)".

input efficiency means that a given level of outputs cannot be produced with less inputs used.

The efficiency can be measured using two perspectives: with and without taking into account the prices of inputs and outputs. The technical efficiency focuses purely on the increase of output or decrease of input to reach the theoretical production frontier. The allocative efficiency, on the other hand, uses prices of inputs and outputs in order to minimize the cost or maximize the revenue. The target of the allocative efficiency is not only to use no excess amount of inputs above the minimal level needed (given the level of outputs), but to find the cheapest combination of inputs that can achieve it – or from the output efficiency perspective, not only to produce the maximal amounts of outputs but to find the most profitable combination.

Figure 2.1: Technical and allocative efficiency



Source: Adapted by the author from Murillo-Zamorano (2004).

The two concepts may be illustrated in figure 2.1 on an example with two inputs and one output. On each axis there is the amount of one of the inputs used per the amount of output produced. The unit isoquant, curve YY', which forms the border of the technological set, shows the minimal combination of inputs x_1 and x_2 needed to produce one unit of output. All points inside the technological set, such as point P, are therefore inefficient. The level of technical efficiency is estimated by the relative distance between the origin and the projection of the observation on the unit isoquant (OR) as compared to the whole distance of the observation from the origin (OP). It is measured by the ratio $TE = OR/OP$.

To find the level of allocative efficiency we need the input prices to form the isocost-line CC', which depicts the sets of inputs of the same total price. The allocative efficiency is the relative distance, from the origin, of projection of point P on the isocost-line (OS) to the projection of point P on the unit isoquant (OR). It may be expressed as $AE = OS/OR$. From the technical and allocative efficiency the whole economic efficiency (which Farell originally denoted as the overall efficiency) may be computed by $EE = TE \times AE = (OR/OP) \times (OS/OR) = OS/OP$. Among all the technically efficient points on the unit isoquant, only the point R' is allocatively efficient as well. The allocative efficiency is thus far more restrictive than the technical efficiency but it also requires deeper insight in the analysed field and the knowledge of input prices.

Efficiency as a concept is closely related to effectiveness. While the efficiency reflects the process of creation of goods or services, the effectiveness considers the use of these goods or services for the achievement of objectives to which they are designed, i.e. the outcomes. It thus measures the optimal production of output to achieve the outcome. The effectiveness is however more complicated to assess, especially in the public sector. There may be multiple objectives that can influence each other rendering the analysis too complicated. The outcomes may be under the influence of current political representation. Thus most of the analyses of public spending concentrate only on the efficiency.

The efficiency and effectiveness analysis of the public sector offers several specific issues. Public organisations are typically owned by the general public. This leads to lower level of monitoring and lower incentives to search for optimal solution. It is also not easy to define appropriately the inputs and outputs, not to mention the objectives for the public spending analyses. Even when there is an agreement about the definitions, indicators reflecting the quantity and not quality of the service are usually available. This is one of the most important limitations of public sector evaluations.

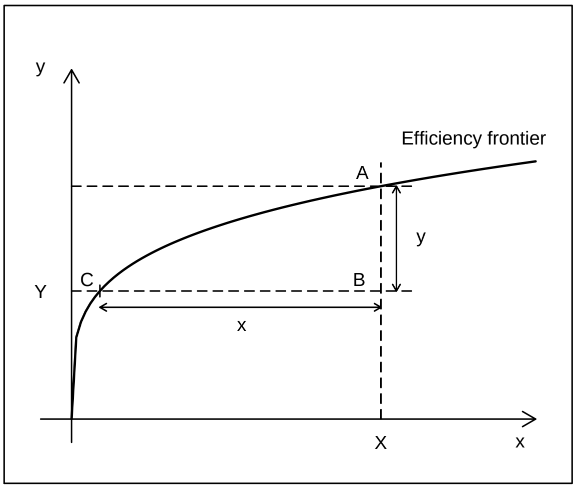
Furthermore, the estimation of allocative efficiency may be impossible because the public services are often not traded on the market. Thus the competitive prices of public services are usually unknown. For these reasons most of the public spending analyses, including the current analysis, focus only on the technical efficiency. More about efficiency analyses of public sector can be found among others in Kalb (2010).

2.2 Different ways to measure public spending efficiency

The efficiency analysis is based on the assumption that all DMUs have the same production function therefore each efficient DMU can produce the same amount of output if it employs a given amount of input. To measure the efficiency we then need to relate the observed performance of a DMU to the maximal performance possible for each combination of inputs – the so called productivity possibility frontier (PPF). While the optimum may never be known in reality, we can use the data to estimate it based on the fact that the optimal production cannot be worse than the production of the best performing unit observed.

Once the PPF is estimated, we can use it to infer efficiency as illustrated in figure 2.2 for a simple single input - single output analysis. Comparing units A and B, or C and B, we can see that the unit B is clearly inefficient. From the point of view of the output efficiency the unit B could use the same amount of input but increase the output produced by y to reach the level of unit A. If we consider the input efficiency, the unit B could decrease the input used by x while maintaining the same output produced to become as efficient as unit C. Units A and C are assumed to lie on the PPF (here called efficiency frontier) for the lack of further knowledge about the frontier.

Figure 2.2: Input and output efficiency



Source: Adapted by the author from Mandl *et al.* (2008).

The PPF may be estimated using various methods. These may be divided based on several criteria: 1) the need for ex-ante functional form assumption and 2) the way the errors are viewed. The first criterion divides the methods into two groups, the non-parametric and the parametric methods, where the parametric ones need assumptions on the production function. Based on the second criterion, the methods may be divided into deterministic and stochastic. All these estimators are originally defined for the cross-sectional data but they are adjusted to manage panel data as well.

Deterministic methods are based on the assumption that all deviations from the PPF are caused by factors that may be influenced by the DMU and thus are considered as inefficiencies. Therefore it does not take into account exogenous factors such as the environment or various sociological factors. This disadvantage is improved by the stochastic methods which can model the frontier with a stochastic error (by creating a composite error).

The parametric methods assume a particular functional form of the production function and use econometric techniques to estimate parameters of the functional form based on the data. Depending on the technique used, the function is shifted so that only few or no data points lie above the function. The most often used parametric method is the Stochastic Frontier Analysis (SFA). The Deterministic Frontier Analysis, which may be also performed by linear programming, is used mainly for comparison as it does not offer any additional advantage.

The traditional non-parametric methods are deterministic. The Data Envelopment Analysis (DEA) and the Free Disposal Hull (FDH) are the most often used methods in this category. These methods use mathematical programming to find an upper hull that envelopes all the data points. It is technically easier to apply these methods to datasets with higher number of inputs and outputs (even though they may require higher number of observations to perform well). The difference between DEA and FDH lies in the assumption of convexity in DEA. FDH compares each observation only with real DMUs not with a convex combination of real DMUs, which renders the comparison more realistic but increases significantly the number of units FDH classifies as efficient.

To account for the limitations of the most commonly used non-parametric methods new approaches have been developed. Partial Frontier Analysis (PFA) seeks the answer to the issue of sensitivity of DEA and FDH to outliers while not assuming a particular production function. Furthermore, the deterministic feature of DEA and FDH limiting the analysis lead to the introduction of a stochastic non-parametric method, called the Stochastic Nonparametric Envelopment of Data (StoNED).

Chapter 3

Non-parametric deterministic methods

This chapter aims to analyse and compare the deterministic non-parametric methods briefly introduced in the previous chapter. The description of the methods follows their historical development with the DEA first and the FDH, developed in response to DEA, second. The third method, the SFA, which was introduced as the solution to the issue of the deterministic characteristic of the previous methods, is presented in the next chapter along with the most recent method, non-parametric stochastic approach, developed to improve shortcomings of the previous.

3.1 DEA

The Data Envelopment Analysis is the first and most commonly used non-parametric efficiency estimator. It has developed from the approach introduced by Farrell (1957). Since then it has been considerably analysed and extended. In this section we analyse the origin, the set-up and the use of the method to enable further comparison with the other methods.

3.1.1 Literature review

To set the DEA in the historical context let us shortly recount the early evolution of efficiency measures. The first attempts to measure efficiency included various indices and performance indicators. These can however measure only performance as they are not able to compare the current production to the optimum. The evolution of the currently used efficiency measures starts with

Farrell (1957), who is the most cited author in the field of efficiency measurement.

Farrell (1957) is building on the ideas from Koopmans (1951) and Debreu (1951). Koopmans (1951) introduces the notion of technical efficiency. According to Koopmans (1951) a production process is efficient if it cannot produce more of any output from given amount of input (other outputs fixed) or cannot use less amount of any input in the production of given output (other inputs fixed). However, this measure cannot be evaluated in reality as the maximal feasible production of a DMU is often unknown and thus the optimal frontier needs to be estimated. Debreu (1951) provides a radial measure of the technical efficiency.[2] Farrell (1957) uses a modified version of the Koopmans' definition, comparing the efficiency relative to the best performing units, and creates a measure to study the overall efficiency, composed of technical and allocative efficiency. He reasons that in order to reach the optimum, producer may need not only to change the scale of inputs or outputs but also to choose the optimal composition of input and output.

The Farrell's contribution was further enhanced by (Charnes *et al.* 1978, 1981) by employing the linear programming method. They named the resulting approach the Data Envelopment Analysis. Both Farrell (1957) and Charnes *et al.* (1978) assume the convexity of the technological feasibility set, strong disposability of inputs and outputs[3] and constant returns to scale (CRS). As the assumption of CRS is often not satisfied, the use of variable returns to scale (VRS) is implemented among others by Banker *et al.* (1984). Thorough summary of the DEA bibliography may be found in Cooper *et al.* (2007).

Recent development of the DEA method seeks to improve on the limitations of the deterministic non-parametric methods. This includes bootstrapping and asymptotic results to enable statistical inference. It also leads to the introduction of new robust method which is described later in this thesis, the Partial Frontier Analysis (section 3.4).

[2]Radial measure searches for optimum using fixed composition of inputs and outputs and scaling down inputs or scaling up outputs equiproportionally.

[3]Strong disposability, in other fields also called free disposability, distinguishes from a weak disposability assumption in the notion that an increase of strongly disposable inputs cannot decrease outputs produced.

3.1.2 Description of the estimator

As noted in the previous subsection, the DEA is building on the initial measure of Farrell (1957) and therefore it assumes that all DMUs have identical production function, convex technological feasibility set and strong disposability of inputs and outputs. We first introduce the estimator using CRS and then describe the extension to VRS. The DEA method is presented here only for technical input efficiency. The specification of the method for output efficiency and introduction of additional concepts, such as allocative efficiency, may be found in Coelli *et al.* (2005) or Cooper *et al.* (2011).

We have I firms producing N outputs out of M inputs. We denote X the $M \times I$ input matrix (with columns x_i being the inputs of firm i) and Y the $N \times I$ output matrix (with columns y_i being the outputs of firm i).

The DEA may be expressed using various optimization problems. The first is the ratio form introduced by Charnes *et al.* (1978). For each firm i it computes the maximal ratio of weighted outputs and inputs $\frac{w'y_i}{z'x_i}$ by choosing w and z, the output and input weights (with dimensions $N \times 1$ and $M \times 1$ respectively), subject to the condition that the ratio for the other firms using the same weights is smaller or equal to one and the weights are non-negative. Thus a firm is inefficient if, even using the most favourable weights, its ratio of weighted outputs to weighted inputs is lower than one.

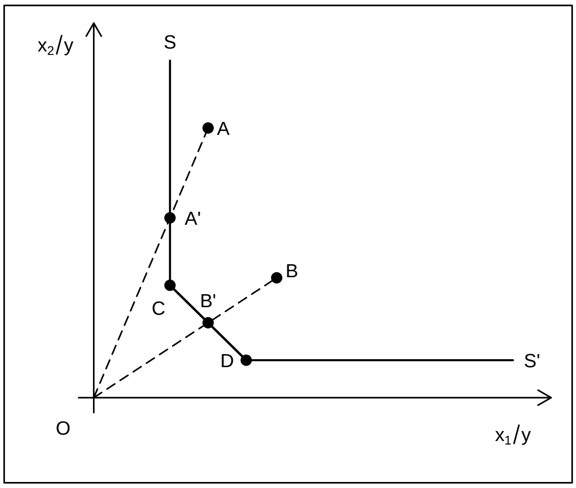
$$
\begin{aligned}
\max_{w,z} \quad & \frac{w'y_i}{z'x_i} \\
\text{subject to} \quad & w'y_j/z'x_j \leq 1 \quad \text{for} \quad j = 1, 2, \ldots, I \\
& w, z \geq 0.
\end{aligned}
\tag{3.1}
$$

However, as pointed out by Coelli *et al.* (2005), the ratio form has infinite number of solutions (if (w^*, z^*) solve the optimization problem, then (aw^*, az^*) does too). This can be solved by imposing further restriction, $z'x_i = 1$, creating thus the multiplier form. From the multiplier form an equivalent form, called the envelopment form, may be derived using the duality in linear programming. This is the most commonly applied form of the DEA. In this representation λ_i is a $I \times 1$ vector of firms' weights in the optimization for firm i, θ_i is the efficiency score of firm i and the rest stays as defined previously.

$$\begin{aligned}
\min_{\theta_i,\lambda_i} \quad & \theta_i \\
\text{subject to} \quad & -y_i + Y\lambda_i \geq 0 \\
& \theta_i x_i - X\lambda_i \geq 0 \\
& \lambda_i \geq 0.
\end{aligned} \tag{3.2}$$

This optimization problem examines every linear combination[4] of output vectors $Y\lambda_i$ which is larger or equal to the output vector of firm i, y_i. Using the same weights, the corresponding combination of input vectors $X\lambda_i$ is then compared to the input vector of firm i, x_i, to find the lowest ratio of the input which is still larger or equal to the combination of inputs. This can be understood as a comparison to hypothetical firm constructed as linear combination of all the firms in the sample. For an efficient firm, the only ratio of its inputs which may be at most equal to the hypothetical firm's inputs, while the outputs of the hypothetical firm are larger than the efficient firm's outputs, is equal to one. Thus if firm is efficient, there exist no linear combination of all the firms that would produce more outputs with less inputs.

Figure 3.1: Input and output slacks



Source: Adapted by the author from Coelli *et al.* (2005).

When the DEA method is applied, the production possibility frontier is constructed as piece-wise linear which is caused by the assumptions of convexity and strong disposability of inputs and outputs. Yet if only efficiency scores are

[4]For CRS the linear combination does not have to be convex. This changes for VRS specification.

reported, this form may cause imprecisions in the efficiency analysis because of the sections that are parallel with the axes as can be seen in figure 3.1.

In this example with two inputs and one output the efficiency frontier is determined by the firms C and D and the efficiency of firms A and B are estimated as OA'/OA and OB'/OB, respectively. The firm A' is constructed as efficient because the inputs needed for production of one output cannot be decreased for a fixed proportion of inputs. However the point A' is not efficient in the sense of non-radial efficiency since the same amount of output could be produced using lower level of input x_2. The overused input can be at most decreased by $A'C$. This amount is known as the input slack. In case of more outputs, similar problem could arise in the output dimension leading to output slack which could be defined correspondingly.

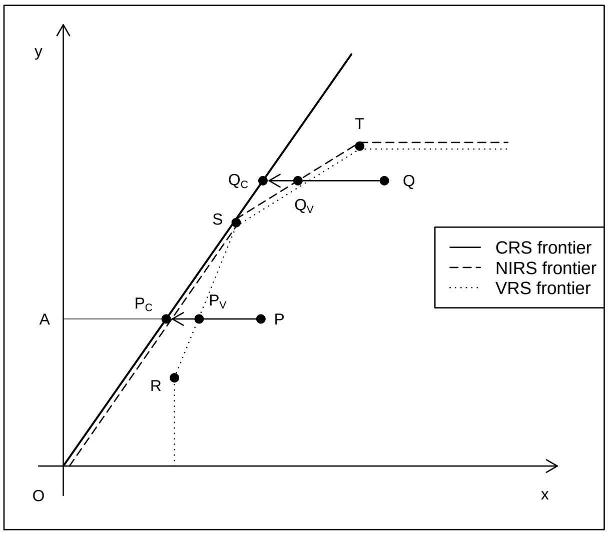
The slacks may be estimated using additional linear programming problem (see Coelli *et al.* 2005). It is recommended in the literature to report not only the efficiency values but also the input and output slacks. Some authors (for example Cooper *et al.* 2011) call units with efficiency score equal to one but positive input or output slacks as weakly efficient. However, slacks may be seen as more important than they actually are. Their existence depends on the choice of estimator and with increasing sample size their significance decreases.

3.1.3 The expansion of the method for VRS

Until now we assumed the constant returns to scale (CRS) which means that firms can increase the production α-times by scaling all inputs by the same constant (for one output $f(\alpha x) = \alpha f(x), \alpha > 0$). Yet in practice it can both happen that increasing all inputs equiproportionally increases the outputs in higher or in lower proportion. This is illustrated in the figure 3.2.

This graph compares frontiers and efficiency scores based on three assumptions on returns to scale: constant returns to scale (CRS), variable returns to scale (VRS) and non-increasing returns to scale (NIRS) that partially combines the previous two. As can be seen from the points P and Q the efficiency scores varies with the assumption. Based on the CRS assumption only S and the theoretically constructed point P_C are efficient. Replacing the assumption by VRS firms R and T become efficient as well. P and Q are still inefficient but their scores increase, for point P from AP_C/AP to AP_V/AP. To determine the influence of the scale factor on the "total" technical efficiency (assuming CRS)

Figure 3.2: Frontiers with different returns to scale assumptions



Source: Adapted by the author from Coelli *et al.* (2005).

it may be decomposed into "pure" technical efficiency (under the assumption of VRS) and scale efficiency (SE).

$$\theta_{\text{CRS}} = \frac{AP_C}{AP} = \frac{AP_V}{AP} \times \frac{AP_C}{AP_V} = \theta_{\text{VRS}} \times \theta_{\text{SE}} \tag{3.3}$$

To account for VRS mathematically, convexity constraint $1_I\lambda_i = 1$ is added to the linear programming problem (3.2), where 1_I denotes an $I \times 1$ vector of ones. Using this the linear combinations are convex and firms are compared only with other firms on equivalent production level. To determine whether firm operates under increasing or decreasing returns to scale, an additional linear programming problem can be computed under the assumption of NIRS which is implemented by changing the convexity constraint to $1_I\lambda_i \leq 1$. For the resulting efficiency scores the returns to scale are determined using the following set of inequalities.

$$\begin{aligned}
\theta_{\text{VRS}} > \theta_{\text{NIRS}} = \theta_{\text{CRS}} &\rightarrow \text{increasing returns to scale}\\
\theta_{\text{VRS}} = \theta_{\text{NIRS}} > \theta_{\text{CRS}} &\rightarrow \text{decreasing returns to scale}\\
\theta_{\text{VRS}} = \theta_{\text{NIRS}} = \theta_{\text{CRS}} &\rightarrow \text{constant returns to scale}
\end{aligned}$$

3.1.4 Further extensions

The generalization of DEA did not stop with dropping the CRS assumption. One of the further contributions to the basic DEA estimator is the inclusion of non-discretionary inputs and outputs because firms typically do not have control over all variables. Among the non-discretionary variables belong variables fixed in short-run (for example number of factories) or some exogenous variables which are observed and therefore can be directly included in the analysis (as examples of these inputs rainfall or water pollution could be named). Banker & Morey (1986) incorporate this extension into the linear programming problem (3.2) by dividing inputs into 2 groups where only the flexible variables under control are expected to be adjustable by firms. This appears in the linear programming problem as two separate inequalities for the inputs, with the efficiency score figuring only in the inequality for discretionary inputs (X_D). Similar adjustment could be done in case of non-discretionary outputs (Y separated to Y_D and Y_N).

$$\begin{aligned} \theta_i x_{Di} - X_D\lambda_i &\geq 0 \\ x_{Ni} - X_N\lambda_i &\geq 0 \end{aligned} \tag{3.4}$$

In equation (3.4) both discretionary and non-discretionary inputs have a positive (or at least non-negative) effect on outputs. This assumption can be further relaxed. In the beginning of this analysis we set the assumption of strong disposability which excluded the case of input congestion. However the example of water pollution demonstrates the need to account separately for inputs contributing positively and negatively to the production. This leads to further division of the variables according to the sign of their contribution. The reason for input congestion lies often in non-discretion. In that case the optimization problem is adjusted by adding further inequality with the reverse sign for the “undesired” inputs, as discussed in Ray & Chen (2015). A more general estimator including inputs with negative marginal effect is specified in Coelli *et al.* (2005) or Färe *et al.* (1985). Analogically “undesired” outputs could be included using separate inequalities for outputs.

3.1.5 Panel data in DEA

When panel data is available, more insight can be gained on the change of efficiency over time. The simplest way is to pool the data, treating all observations as separate DMUs. Pooling is not optimal because it assumes that each observation is an independent yet comparable unit and it does not take into account overall change of technology. Slightly advanced technique to detect efficiency trends of units is the window analysis. It takes k adjacent periods and pools data only from these periods. The use of rolling window in order to cover all t periods produces $(t-k+1)$ separate DEA analyses. The advantage of this approach is that it does not expect the technology to remain the same over the whole studied period. At the same time restricting the analysis only to one period can make the analysis infeasible when there are only few units and many input and output variables. The length of the window, k, is chosen to balance this trade-off. This approach however still needs the assumption of independent observations and constant technology over each "window" of data. In addition to this, the observations from the first and last periods are analysed much less than the rest of the sample putting unequal weights on different observations.

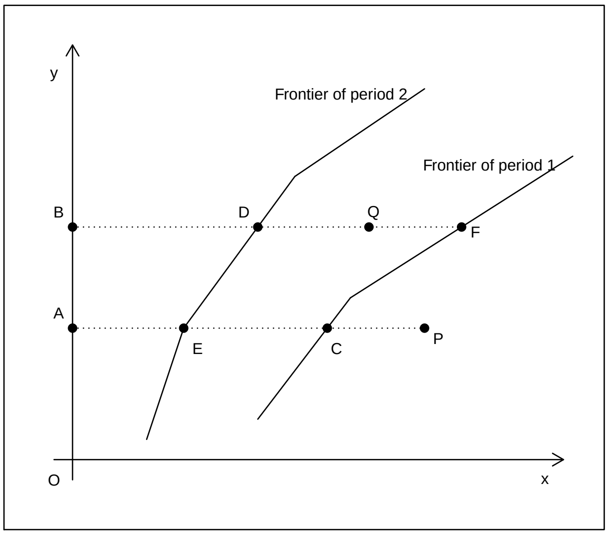
The method using Malmquist index solves some of these issues. It takes into account not only relative efficiency but also the over-period changes of the production frontier. It compares adjacent periods for each decision making unit. The index can be decomposed into two parts, "catching up" with the frontier and "frontier shift". This can be seen on an example in figure 3.3. "Catching up" is calculated as the ratio of efficiency in year 1 to efficiency in year 2. For an input efficiency this can be expressed as $C = \frac{BD/BQ}{AC/AP} = \frac{\theta_2^2}{\theta_1^1}$ where θ_t^s stands for efficiency of observation in year t when compared to efficiency frontier in year s.

The "frontier shift" denotes the ratio of efficiencies comparing each observation to both of the frontiers. The "frontier shift" from the point of first year has the following formula: $\phi_1 = \frac{AC/AP}{AE/AP} = \frac{AC}{AE} = \frac{\theta_1^1}{\theta_1^2}$. For the observation in the second year, the expression is the following: $\phi_2 = \frac{BF/BQ}{BD/BQ} = \frac{BF}{BD} = \frac{\theta_2^1}{\theta_2^2}$. The combined "frontier shift" is formulated as the geometric mean of the ratios for the first and the second year $FS = \sqrt{\phi_1\phi_2}$.

The resulting Malmquist index is defined as the product of the "frontier shift" and "catching-up" scores: $MI = C \times FS = \frac{AP}{BQ}\sqrt{\frac{BD}{AC}\frac{BF}{AE}} = \sqrt{\frac{\theta_2^1}{\theta_1^1}\frac{\theta_2^2}{\theta_1^2}}$. For the Malmquist index it holds that the resulting ratio bigger than one means a positive improvement - either increase in relative efficiency (when "catching up" is bigger than 1) or technological innovation (when "frontier shift" is bigger than one) or possibly both.

Outside this simple example, Malmquist index is calculated using the ef-

Figure 3.3: Malmquist index: catching up and frontier shift



Source: Adapted by the author from Cooper *et al.* (2007).

ficiency scores from DEA analyses applying appropriate form of the problem. For analyses with more than 2 time periods, each two adjacent periods are analysed to produce a score of the Malmquist index. For input oriented analysis with constant returns to scales, the form in equation (3.2) is used. When variable returns to scale are assumed, the convexity constraint $1_I\lambda_i = 1$ is added to equation (3.2). Literature also offers alternatives accounting for slacks or analysis combining input and output orientation.

3.2 FDH

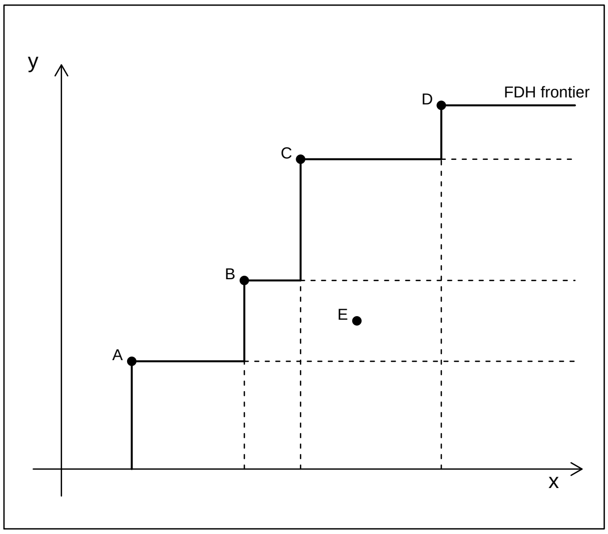
The Free Disposal Hull (FDH) estimator is the second deterministic non-parametric method for efficiency measuring. It has developed from DEA. The difference between DEA and FDH lies in the assumptions on the technological feasibility set. While DEA assumes convexity and strong disposability, FDH maintains only the disposability assumption. This implies that FDH compares each unit only with other existing units, not with a linear combination of existing units. While this causes certain limitations (discussed later), it increases comparability of the scores. The origin of this method dates back to Deprins *et al.* (1984) that introduce FDH as an alternative to convex approach based on Farrell (1957).

3.2.1 Description of the method

The FDH estimator compares all input and output levels in order to detect inefficiencies. It concludes that firm i is inefficient if it is dominated by firm j in the sense that firm j produces at least as much of all outputs using the same amount or less of all inputs and strictly smaller amount of at least one input (or strictly higher amount of at least one output):

$$y_j \geq y_i \wedge x_j \leq x_i \wedge (x_j \neq x_i \vee y_j \neq y_i) \tag{3.5}$$

Figure 3.4: Free Disposal Hull - Illustration



Source: Adapted by the author from De Borger & Kerstens (1996).

This is illustrated in figure 3.4 showing the graphical representation of the FDH frontier. Units A–D are efficient, unit E is inefficient. The method introduces the notion of “efficiency by default” for units that do not dominate any other unit. These units are not entirely comparable to the other efficient units because the efficiency may be caused purely by the lack of observations on similar production level. This causes the sparsity bias which means that in small samples large fraction of the units is clasified as efficient (see Vanden Eeckhaut *et al.* 1993). The example of units that are efficient by default are units A and D in figure 3.4.

The FDH may be estimated applying the same linear programming problem as for DEA with one difference in the weights. As FDH does not allow for the convex combination of DMUs, the weights can only be equal to 0 or 1. Along with the assumption that the sum of weights equals to one, this implies that each unit is compared only with one other existing unit.

$$\begin{aligned}
\min_{\theta_i,\lambda_i} \quad & \theta_i \\
\text{subject to} \quad & -y_i + Y\lambda_i \geq 0 \\
& \theta_i x_i - X\lambda_i \geq 0 \\
& \lambda_{ij} \in \{0,1\} \quad \text{for} \quad j = 1,...,I, \quad \sum_{j=1}^{I} \lambda_{ij} = 1.
\end{aligned} \tag{3.6}$$

Same as for the DEA, we provide only the input efficiency FDH measure. The output efficiency measure may be found, among other, in De Borger *et al.* (1994), who also introduce a third perspective to measure efficiency, the graph measure. The graph measure combines the input and output efficiency measures and produces thus a single set of scores. In comparison to common approach in literature (measuring either input or output efficiency), De Borger *et al.* (1994) claim that the graph measure generalizes the analysis, especially when it is unclear whether a unit has higher influence over its inputs or outputs. It also solves the problem of ambiguity that may occur when both input and output efficiency is computed and contrasted.

When panel data are available for the analysis, FDH may use the window analysis approach described in section 3.1.5. Same as for DEA, this approach assumes independence of observations in each "window" of data but allows the technology not to stay constant over the whole examined time period. Thus it allows the comparison to DEA even when panel data analysis is performed.

3.2.2 Comparison to DEA

As was mentioned before, FDH method developed from the original DEA relaxing the requirement of convexity. It thus belongs to the methods with the least assumptions imposed (the list of the axioms defining FDH may be found in De Borger *et al.* 1994). The low number of assumptions results in several advantages and disadvantages.

One of the advantages of FDH is that units are compared only with real units. Therefore the comparison is more valuable in the sense that the dominating units provide a real-world example that the efficiency can truly be achieved. This may be violated in case of CRS DEA, which allows for comparison of productions that have completely different scales. The DEA using VRS assumption provides more realistic comparison as the convex combination of units has to lie between the units that determine it. Thus for reasonably large samples the linear combination has very similar scale of production.

According to De Borger & Kerstens (1996) small number of requirements for the FDH method also suits the public efficiency analyses. The reason lies in the fact that economists still do not agree in general on the way to model government behaviour. FDH is therefore often used to analyse public organizations.

On the other hand, one of the disadvantages of the small number of assumptions is the bias caused by the "efficiency by default", mentioned in the previous subsection. This problem may occur in DEA with VRS as well. Yet FDH approach to data classifies large percentage of units as efficient also for the reason that the frontier is not convex (see figure 3.4, unit B in particular). Therefore the contribution of FDH may lie rather in ability "to identify the most obvious cases of inefficiency, rather than to characterize efficiency in itself" (Vanden Eeckhaut *et al.* 1993, p. 308).

3.3 Statistical inference

The previous section investigated the advantages and disadvantages of DEA and FDH methods when compared to each other. While praised for low number of assumptions, these methods share some disadvantages as well. Apart from the deterministic characteristic, which does not allow for stochastic errors, the most important disadvantage of the deterministic approaches is the impossibility of statistical inference when the results are estimated using only linear programming. However, there has been a significant progress in the solution of this problem. The deterministic methods are partially incorporated in statistical framework and there are procedures formulated to obtain convergence rate, confidence intervals, and bias of the efficiency estimators.[5]

To be able to incorporate the deterministic efficiency estimators into statistical framework, several assumptions are needed. The sample is assumed to be a realization of an independent identically distributed random variable that has probability density function $f(x, y)$ with a bounded support Ψ. The probability of observing a unit in an open neighbourhood of the frontier (boundary

[5]For additional information, the reader is referred to Simar & Wilson (2011).

of Ψ) needs to be strictly positive. Finally, for technical reasons the boundary needs to be sufficiently smooth.

It has been proved that both DEA and FDH estimators are consistent with convergence rate that decreases with higher number of inputs and outputs (p and q are the number of inputs and outputs respectively). The original formulation and further information for DEA can be found in Kneip *et al.* (1998), and for FDH in Park *et al.* (2000). As can be seen in equation (3.7), the DEA estimator converges at the rate of $n^{\frac{2}{p+q+1}}$ and FDH estimator at the rate $n^{\frac{1}{p+q}}$. In the equation (3.7) $\theta(x, y)$ refers to the true efficiency and $\hat{\theta}_{DEA}(x, y)$ and $\hat{\theta}_{FDH}(x, y)$ to the respective estimated score using DEA ad FDH estimators. The dependence of the convergence rate on sum of the number of inputs and the number of outputs $(p+q)$ illustrates the curse of dimensionality for the deterministic estimators. That means that although any number of input and output variable is theoretically admissible, with higher number of variables the estimators perform significantly worse.

$$\begin{aligned} \hat{\theta}_{DEA}(x, y) - \theta(x, y) &= O_p(n^{\frac{2}{p+q+1}}) \\ \hat{\theta}_{FDH}(x, y) - \theta(x, y) &= O_p(n^{\frac{1}{p+q}}) \end{aligned} \tag{3.7}$$

While the asymptotic results are theoretically important, they are not always available. Furthermore, the asymptotic feature introduces additional noise in the estimation. Therefore it is useful to study the bootstrapping techniques as well. Bootstrapping is a simulation method which uses the data sample to examine the properties of an estimator. It can be employed to estimate variance and confidence interval of point estimators. The process uses repeated sampling with replacement. For the efficiency frontier, there are two ways to sample: using the observed points and using estimated efficiency scores expressed as distance functions.[6] When the estimated distance is used, the procedure of a simple bootstrap is the following.

Given data, estimate the data generating process ($\widehat{DGP}$) and the distance function $\widehat{\delta}(x, y)$. Use the estimated results to generate B samples. For each sample estimate the efficiency score using DEA method and transform the score to the distance function $\widehat{\delta_b^*}(x, y)$ for $b = 1, ..., B$. The results form an empirical distribution of the distance function $\widehat{\delta^*}(x, y)$. In case of consistency of the

[6]The Farell's efficiency scores are based on the mathematical concept of distance functions by measuring the ratios of actual and optimal input or output. For some derivations it is more convenient to express the scores in terms of a ratio which for output efficiency it attains values in $(0, 1]$ and for input efficiency values in $[1, \infty)$. These are referred to as output and input distance functions.

bootstrap method, this empirical distribution can be used to make statistical inference about the unknown true distribution of the distance function $\delta(x, y)$, given true data generating process (DGP).

$$(\widehat{\delta}^*(x,y) - \widehat{\delta}(x,y)|\widehat{DGP}) \overset{approx.}{\sim} (\widehat{\delta}(x,y) - \delta(x,y)|DGP) \tag{3.8}$$

The consistency of the bootstrap method depends on the re-sampling process. The naive bootstrap assigns each of the n observed values probability $1/n$ and samples randomly n draws from this distribution. However when the observed dataset comes from bounded support, as we assumed for the efficiency frontier analysis, the estimates are inconsistent.[7] This may be solved using sub-sampling, which means drawing samples smaller than the total number of observations, or using smoothing techniques. Smoothing techniques use a smooth estimate of the distribution of the data instead of the empirical distribution. They apply the kernel estimator. To ensure consistency of the smooth techniques even around the boundary, reflection method is used which creates a reflection of the data through the frontier. A simple version of the smoothing techniques is the homogeneous smooth bootstrap, formulated in Simar & Wilson (1998). It needs a further assumption: the inefficiencies are distributed homogeneously over the input-output space.

Further properties include the bias of the estimators. Even though it is consistent, the DEA estimator produces biased estimates. It can be shown that the estimated efficiency is higher than the true efficiency because we construct the frontier using only observed units and not the theoretically most optimal. The efficiency scores produced by FDH may be even more biased due to the non-convexity of the efficiency frontier. Using the mean of the bootstrap distribution, it is possible to correct for the bias. However, such correction introduces additional noise and thus is usually not recommended, unless the bias is large in comparison to the standard deviation.

3.4 PFA

As already pointed out, the non-parametric methods, DEA and FDH, have several disadvantages. To solve the issue of sensitivity to extreme values and

[7] There is a non-zero probability that the efficient units are sampled and the distance to frontier stays the same. Thus the inference is inconsistent because this probability does not disappear with increasing n.

outlying points of non-parametric methods, partial frontiers estimators have been established.[8] Partial frontiers estimators are based on FDH estimation but they do not envelop all the sample data. Some observations are allowed to lie above the efficiency frontier and thus the partial frontier estimators produce more robust results. There are two types of partial frontier approaches, the order-m efficiency and order-α efficiency.

The order-m estimator performs the FDH analysis including only a random draw of m peer firms. For input efficiency it takes the population of observed firms that produce at least the same amount of output as the studied unit (these are called the peer firms), repeatedly selects m units from the specified population and performs the FDH analysis. Thus a firm's input level is not compared to the absolute minimum among those who produce at least at the same output but rather to the expected minimum input level from a draw of m random peer firms. If the firm is more efficient than a random draw of m peers, the order-m efficiency score can be greater than one. This cannot happen for classical non-parametric estimators as those always include each unit as their own "peer". The parameter m is a chosen fixed number which can either be seen as the trimming parameter, serving for robustness, or as a parameter for benchmarking in peer analysis. In limit as $m \to \infty$ the order-m estimator converges to the FDH, the so called full frontier estimator.

The formula for the order-m efficiency is based on FDH as may be seen in equation (3.9). The investigated unit uses x inputs and produces y output. The efficiency score is estimated to be the average of the lowest fractions of the inputs x that are still higher or equal to X_i^d, that means the average of D FDH efficiency scores. X_i^d stands for the input level of one of the firms from the d-th repeated draw that produce at least the same amount of output as y.

$$\theta_m(x,y) = \frac{1}{D}\sum_{i=1}^{D}\theta_m^d, \text{where} \quad \theta_m^d = inf\{\theta|\theta x \geq X_i^d, i = 1,...,m\} \tag{3.9}$$

For order-m frontier analysis, m can be manipulated to change the fraction of firms above the efficiency frontier. On the other hand, the order-α frontier uses a reverse approach. It chooses the probability $100-\alpha$ that a firm lies above the α-frontier. The parameter α may be interpreted in an appealing way: if a firm has $\theta_\alpha(x,y) = 1$, there is $100 - \alpha$ probability of another firm being

[8] For more detailed information the reader is referred to Daraio & Simar (2007).

more efficient, that means to be above the order-α frontier. Thus a firm with $\theta_\alpha(x, y) = 1$ may be called efficient at level $\alpha\%$. A firm with order-α efficiency score greater than one is seen as super-efficient. Thus firms can be compared to α-quantile of firms using the frontier. The order-α frontier coincides with the FDH frontier when $\alpha = 100$.

To estimate the order-α efficiency we need to formulate probabilities based on the production process. We investigate again a firm which produces y output out of x inputs. The probability of this firm being dominated is studied as the probability that other firms have lower inputs conditional on their output being higher that y. Conditional on this probability being higher than $100-\alpha$, efficiency is estimated as the lowest fraction of inputs x for which a firm would still be dominated.

$$\theta_\alpha(x, y) = inf\{\theta | Prob(X \leq \theta x | Y \geq y) > 100 - \alpha\} \tag{3.10}$$

Generally, the PFA methods produce more robust results because outliers and other unexpectedly efficient firms are allowed to lie above the frontier and do not shift the frontier. When outliers are present, the classical non-parametric methods, FDH and DEA, may produce (downward) biased efficiency scores because the frontier may shift significantly to include all the data.

However, the partial frontier estimators are based on the FDH method and therefore may suffer from non-smooth frontier and other problematic aspects inherited from the original estimator. Furthermore, as the DEA and FDH these estimators may be judged by the deterministic characteristics. In response to the criticism of deterministic estimators the stochastic counterparts have been introduced. These are presented in the next chapter.

Chapter 4

Parametric and recent methods

The previous chapter describes two classical non-parametric deterministic methods for efficiency measurement and a recent robust estimator, the partial frontier analysis. These methods do not require any additional parametric assumption. However their deterministic feature causes some drawbacks which resulted in the introduction of a stochastic parametric method, the Stochastic Frontier Analysis. This method is based on econometric approach. It includes a composite error into the frontier estimation to account for the exogenous factors and therefore it requires assumptions on functional relationship.

Nevertheless SFA has some disadvantages as well. It is criticised for the necessity of the assumption of the production function. This leads to an introduction of a new method combining the non-parametric frontier with inclusion of stochastic errors. This method is described in the third part of this chapter (section 4.3). The last section summarizes the methods and comments on the use in this analysis and in general.

4.1 DFA - the precursor to SFA

The SFA was not the first parametric method to be described. It developed from a method which is now called the Deterministic Frontier Analysis (DFA). The parametric approach to efficiency analysis started with Aigner & Chu (1968) who, inspired by Farrell's ideas, tried to estimate the production frontier using Cobb-Douglas production function. As the estimation method they still use the linear programming (offering quadratic programming as an alternative).

The DFA requires functional form assumption but does not include any stochastic error yet. Same as the two previous methods, DEA and FDH, DFA is a deterministic method and thus does not allow for inference when only linear programming is used for the estimation. This leads to attempts to incorporate the ideas behind DFA into econometric models. The estimations thus may be performed using maximum likelihood with different distributional assumptions on the residuals and ordinary least squares (OLS) with some necessary adjustments. The estimation methods are described in the following two subsections.

Even when the econometric estimation methods are applied, DFA is mostly not used in practice as the deterministic features of DFA still cause some problems. Due to possible measurement errors, outliers or exogenous shocks the estimated efficiency scores may be misleading. In reaction to this disadvantage, (Aigner *et al.* 1977) introduce stochastic error in the frontier estimation, reasoning that not every deviation from the frontier may be caused by inefficiency.

4.1.1 Introduction of DFA

Although DFA is not used much in practice, as it does not offer much improvement over the non-parametric methods, we start the description with the DFA and expand it later for the SFA (section 4.2). The setting is introduced as a production frontier estimator for a single output using cross-sectional data. The cost frontier approach is presented in subsection 4.2.2. The extension to panel data is described in section 4.2.3. Production of more outputs could be treated using seemingly unrelated regressions, multiple output distance functions, or by calculating revenues. However this analysis uses only a single output and the extension to more outputs is therefore omitted.

We have I firms and each firm i produces only one output y_i out of M inputs, $x_{i1}, \ldots, x_{iM}$. For the parametric method we need an assumption on the functional form of the relation between the output and the vector of inputs. For the output efficiency estimation the production may be expressed as in the equation (4.1) where y_i is the output of firm i, x_i is the $M \times 1$ vector of inputs of firm i, β is the vector of regression parameters and $TE_i \in (0, 1]$ is the estimated technical efficiency score of firm i. Same as defined theoretically in section 2.1, from the output efficiency perspective the technical efficiency score is expressed by the ratio of output produced y_i to the maximal possible output y_i^*.

$$\begin{aligned} y_i = f(x_i, \beta) \times TE_i, &\qquad \text{for} \quad i = 1, 2, \ldots, I \\ TE_i = \frac{y_i}{y_i^*}, &\quad \text{where} \quad y_i^* = f(x_i, \beta) \end{aligned} \tag{4.1}$$

In order to use simpler linear estimation techniques the function $f(.)$ is typically assumed to be linear or log-linear. Following Murillo-Zamorano (2004) the estimation is described in this study for the Cobb-Douglas function which is log-linear. For $i = 1, 2, \ldots, I$ we have

$$ln(y_i) = lnf(x_i, \beta) + ln(TE_i) = \beta_0 + \sum_{k=1}^{M} \beta_k ln(x_{ik}) - u_i \tag{4.2}$$

In equation (4.2) we may see the production equation using the assumption of log-linear production function and the transformation of the output efficiency score $TE_i = exp(-u_i)$, where $u_i \in (0, \infty)$ represents the amount by which firm i lags behind the optimum. For an efficient firm, the efficiency score equals one, for other it is between zero and one.

So far this subsection has considered only output efficiency. The reason is that with the usually used assumption of Cobb-Douglas production function, input efficiency can be calculated from the output efficiency scores. This can be seen in equation (4.3) when compared to equation (4.2). The last term in equation (4.3) can be equated to the output inefficiency term, u_i. Therefore for the following subsections concerning production frontier, the description focuses solely on output efficiency.

$$ln(y_i) = lnf(x_i e^{-\eta_i}, \beta) = \beta_0 + \sum_{k=1}^{M} \beta_k ln(x_{ik}) - \eta_i \sum_{k=1}^{M} \beta_k \Rightarrow u_i = \eta_i \sum_{k=1}^{M} \beta_k \tag{4.3}$$

4.1.2 Estimation of DFA

The efficiency score can be estimated using both linear programming as well as the econometric techniques. While linear programming was shortly discussed in the previous two chapters, we need to mention the econometric methods to estimate the equation (4.2). These methods may be divided into two groups: methods that need further assumptions on the distribution of the errors, such as maximum likelihood estimation (MLE) method, and those that do not need these assumptions, such as OLS or mean absolute deviation (MAD).

The OLS and MAD methods are typically used to estimate the average behavior. Thus they need to be adjusted by shifting the frontier so that the u_i

component of the error may be directly translated to the efficiency score.[9] This is possible using the corrected OLS (COLS) and corrected MAD (CMAD).

The correction for COLS and CMAD is based on the fact that $TE_i = exp(-u_i)$ is the efficiency score and is therefore restricted to be $TE_i \in (0, 1]$ with at least one efficiency score equal to one. This translates to a condition on u_i which needs to be inside $[0, \infty)$ with at least one error equal to zero. To satisfy this restriction, the intercept and the residuals are adjusted as may be seen in the equation (4.4).

$$
\begin{gathered}
ln(y_i) = \hat{\beta}_0 + \sum_{k=1}^{M} \hat{\beta}_k ln(x_{ik}) + \hat{e}_i = (\hat{\beta}_0 + \max_j\{\hat{e}_j\}) + \sum_{k=1}^{M} \hat{\beta}_k ln(x_{ik}) + \hat{e}_i - \max_j\{\hat{e}_j\}, \\
\text{where} \quad \widehat{TE}_i = exp(-\hat{u}_i) = exp(\hat{e}_i - \max_j\{\hat{e}_j\}) \in (0, 1]
\end{gathered} \tag{4.4}
$$

For the MLE estimation we need to assume some distribution of the residuals. The distribution has to be non-negative so that u_i may represent the lag behind the efficiency frontier. Suggested distributions include the exponential, half-normal, truncated normal, and gamma distribution. Gamma distribution is usually not used as it increases the complexity of the estimation significantly. MLE estimation allows direct inference without the need for bootstrapping techniques (as described in section (3.3) for the non-parametric methods).

However, DFA is not optimal for the reason that, as explained previously, it requires additional assumptions on functional form but does not perform better as DEA or FDH because of its deterministic characteristic. Therefore we shift our attention to the SFA method.

4.2 SFA

4.2.1 Production frontier analysis

In the SFA model the efficiency score is replaced by the composite error which contains not only the efficiency score (represented by u_i) but also the exogenous stochastic error, v_i. In contrary to the previous equation (4.2), this form can account for the statistical noise and therefore the efficiency scores can be separated from exogenous shocks.

[9] It needs to be assumed that the efficiency frontier is the same as the relation between inputs and output for the average firm, only shifted to the level of the most efficient firms.

$$log(y_i) = \beta_0 + \sum_{k=1}^{M} \beta_k ln(x_{ik}) + v_i - u_i, \quad \text{for} \quad i = 1, 2, \ldots, I \tag{4.5}$$

To estimate the SFA model only econometric methods may be used. Furthermore, the methods without distributional assumptions (such as COLS) do not allow to separate the efficiency and the statistical noise rendering the estimated results useless. Therefore practitioners mostly use the MLE methods. The distributions of both parts of the error term along with their independence need to be assumed. Generally v_i is assumed to follow zero-mean normal distribution. As listed before, there are various assumptions on the distribution of u_i, namely exponential, half-normal, truncated-normal, or gamma distribution.

While for DFA it was sufficient to estimate only the parameters, for the SFA method the more important part of the estimation lies in the identification of the efficiency score because from the first step of estimation we obtain only the composite residual. For the identification of the efficiency score most authors follow Jondrow *et al.* (1982). The paper offers two alternatives to estimate u_i: either the conditional mean, $E(u_i|\epsilon_i)$, or the conditional mode, $M(u_i|\epsilon_i)$, where $\epsilon_i = v_i - u_i$. The formulas for half-normal distribution of u_i are included for illustration. Further specific formulas for different distributions may be found in Kumbhakar *et al.* (2015). As ϵ_i is unknown, the residuals from previous estimation $\hat{\epsilon}_i$ are used. In the equation (4.8) σ_u and σ_v refer to the variances of u_i and v_i, the two independent parts of the composite error.

$$E(u_i|\epsilon_i) = \frac{\sigma_* \phi(\frac{\mu_{*i}}{\sigma_*})}{\Phi(\frac{\mu_{*i}}{\sigma_*})} + \mu_{*i} \tag{4.6}$$

$$M(u_i|\epsilon_i) = \begin{cases} \mu_{*i} & \text{if } \mu_{*i} > 0, \\ 0 & \text{if } \mu_{*i} \leq 0 \end{cases} \tag{4.7}$$

$$\begin{aligned} \text{where} \quad \mu_{*i} &= \tfrac{-\sigma_u \hat{\epsilon}_i}{\sigma_u + \sigma_v} \\ \sigma_* &= \tfrac{\sigma_u \sigma_v}{\sigma_u + \sigma_v} \end{aligned} \tag{4.8}$$

To transform the u_i to get the estimates of efficiency scores the following relation is used: $TE_i = E(exp(-u_i)|\epsilon_i)$.

Instead of identifying the efficiency score, we may first perform a test whether there is some significant inefficiency in the model as otherwise we do not need to search for it and may estimate the model using simpler methods. This may be tested by the test for skewness of OLS residuals or the likelihood ratio test. The test for skewness can be done before the estimation is performed. It is based on the assumptions that v_i is symmetric around zero whereas u_i comes from non-negative distribution. The likelihood ratio test on the other hand needs to be performed after the MLE estimation and tests for significance of relevant parameters using log-likelihood statistics. For the half-normal the null hypothesis of the likelihood ratio test is $H_0 : \sigma_u = 0$.

The so-far-discussed model assumes homoscedasticity of both error terms, u_i and v_i. When at least one of them is heteroscedastic, the estimated efficiency scores will be biased. A proposed solution is to let the variances, σ_u and σ_v, depend each on a set of variables and account for the heteroscedasticity using the following parametrization: $\sigma_{u,i} = exp(z_{u,i}\gamma_u)$ and $\sigma_{v,i} = exp(z_{v,i}\gamma_v)$, where $z_{u,i}$ and $z_{v,i}$ are vectors of explanatory variables including a constant. The exponential function is used to ensure that the resulting variances are positive.

The same parametrization can be used to approach a different issue. Inefficiency can be determined by exogenous factors. These factors may even be the focus of an investigation which aims to find the causes of inefficiency and offer a possible reform. Originally, only the simple two-step estimation was available for this type of analysis. The two-step estimation is performed using classical efficiency analysis as the first step and in the second step regressing the resulting efficiency scores on the potential causes using for example Tobit regression. Yet if the explanatory variables of the two steps, x_i and z_i, are correlated, the first step results in biased efficiency estimates. To solve the issue of correlation instrumental variable estimation, possibly in combination with fixed-effects estimation, may be used. Yet even in case of no correlation between x_i and z_i the two-step procedure is not optimal.[10] When feasible, the parametrization introduced to solve the heteroscedasticity issue should be employed. In case of truncated-normal distributional assumption on the residuals, both parameters of the distribution can be parametrized. This means that apart from the parametrization of σ_u we can also express the mean of the distribution, μ, in terms of the same set of explanatory variables, z_u. This increases flexibility of estimation. However, it must be noted that high complexity can cause problems with convergence during the MLE estimation.

[10] When the dependency of inefficiency on a set of explanatory variables is not taken into account, the resulting efficiency scores may have lower dispersion, which can lead to a downward bias of the results from the second-step Tobit regression.

4.2.2 Cost frontier analysis

In the previous subsections the SFA method is described as production frontier estimator. While it may be useful for some analyses, it allows us to study only the technical relation between inputs and outputs. To investigate the economic behaviour, the cost frontier estimation is used. In particular, it may be applied to find how effectively firms minimize their cost given certain amount of outputs produced. Furthermore, it gives us the means to study whether it is possible to decrease costs while maintaining the outputs produced or services provided. In cost frontier estimation the efficiency is mostly studied from the perspective of input efficiency because outputs are assumed to be fixed.

The cost is modelled as the deviation from the optimal cost frontier $C^*(w, y)$, where w are input prices and y outputs. The cost frontier is typically assumed to have translog (transcendental logarithmic) specification. When only one output is produced and the input prices are not different across units, the model can be defined as in equation (4.9). The actual costs are denoted as C_i^a. The variable η_i describes the inefficiency of input usage. Contrary to u_i, it is included with a positive sign to denote the costs used above the optimal level of costs. As in the production frontier model, v_i is the noise term.

$$\begin{aligned} ln(C_i^a) &= ln(C^*(w_i, y_i|w_i = w)) + v_i + \eta_i \\ &= \beta_0 + \beta_y ln(y_i) + \frac{1}{2}\beta_{yy}(ln(y_i))^2 + v_i + \eta_i \end{aligned} \tag{4.9}$$

Due to the features of cost function, partial derivative of the cost frontier with respect to the output has to be inspected if it satisfies monotonicity condition of output: $\partial ln(C^*(w_i, y_i|w_i = w))/\partial ln(y_i) = \beta_y + \beta_{yy} ln(y_i) \geq 0$.

The estimation of the cost frontier is very similar to the estimation of the production frontier. As before (subsection 4.1.2), the model can be estimated with or without further assumptions on the residuals. When no further assumptions are made, COLS and CMAD may be used. These methods are applied similarly to the production frontier estimator as can be seen from the following equation.

$$\begin{aligned} ln(C_i^a) = ln(\hat{C}^*(w_i, y_i)) + \hat{e}_i = ln(\hat{C}^*(w_i, y_i)) + \min_j\{\hat{e}_j\} + \hat{e}_i - \min_j\{\hat{e}_j\} \\ \Rightarrow \widehat{TE}_i = exp(-\hat{e}_i + \min_j\{\hat{e}_j\}) \end{aligned} \tag{4.10}$$

However, when no assumptions on the residual distribution are made, the method belongs to DFA and thus it may suffer from the problems which are inherent to deterministic methods. Therefore MLE with the same set of possible assumptions on the distribution of residuals (half-normal, truncated-normal, exponential, and gamma distribution) is typically used and the procedure is similar as described for the production efficiency analysis in section 4.2.1.

The cost frontier approach may be easily applied when particular inputs are unknown or in case of a production of multiple outputs. It also partially incorporates the notion of allocative efficiency because both technical and allocative inefficiency increase costs and thus optimal costs do not indicate only optimal amount but also optimal composition of inputs. Therefore the cost frontier approach seems to be appropriate for the public sector efficiency analysis.

4.2.3 SFA for panel data

When panel data are available, the SFA method provides us with further possibilities. While it is possible to pool the observations and apply the methods for cross-sectional data, it is mostly not optimal. Panel data models can account for the individual heterogeneity and can potentially be used to study the temporal behaviour of inefficiency.

The models can be divided based on the assumptions on the behaviour of inefficiency in time. When we think that inefficiency is constant, time-invariant models may be used. These models consist of fixed effects model, estimated by the within estimator, and random effects model, estimated either by GLS or MLE estimator. On the other hand, the models with the assumption of time-varying inefficiency have higher complexity.

One of the most flexible estimators is proposed by Cornwell *et al.* (1990). It includes in the regression dummies for each unit and all cross-products of firm dummies with time and time squared. Thus the inefficiency is allowed to change differently for each decision-making unit. This approach does not need any assumption on the distribution of residuals. Yet it may not be optimal in case of high number of units because a long time period would be needed to ensure sufficient number of observations so that all the coefficients can be estimated.

Further approaches are more parsimonious, yet they include composite error and thus they need distributional assumptions for the residuals. Several alternatives assume u_{it} to follow a specific function with the time variable in argument, for example $u_{it} = u_i G(t) = u_i exp[\gamma(t - t_0)]$ (Kumbhakar & Wang 2005). These models are estimated using MLE. Furthermore, there are models that separate individual heterogeneity from efficiency, models that separate persistent and time-varying inefficiency, and models that do both. These models are estimated using multi-step estimations combining fixed or random effects approaches with the MLE estimation using distributional assumptions for all the error terms. The models applied in this thesis are shortly stated during the presentation of results.

4.3 Stochastic non-parametric analysis

Both groups of estimators, deterministic non-parametric and stochastic parametric, have certain disadvantages. Deterministic estimators assign all the deviations from the frontier solely to inefficiency. On the other hand parametric methods are criticised for the functional assumptions needed for the analysis. This leads to an idea to create an estimator which combines the non-parametric and stochastic features in the attempt to avoid the problems of the preceding methods.

First, a technique, called convex nonparametric least squares (CNLS), is presented as an alternative to OLS. This approach uses the framework of SFA without functional assumptions. The function is specified using axioms on its properties, similar to DEA. Based on this technique, a new efficiency analysis method, the Stochastic Nonparametric Envelopment of Data (StoNED), is introduced. A summary of the theory with further specifications may be found in Kuosmanen *et al.* (2015) and Johnson & Kuosmanen (2015).

The CNLS method models the production using a frontier production function $f(x)$ without particular functional form assumption. The production equation looks similar to the framework of SFA with the difference that the CNLS method is specified here for an additive error. However, in case of multiplicate error the equation (4.11) could be logarithmized to obtain similar expression.

$$y_i = f(x_i) + \epsilon_i = f(x_i) + v_i - u_i \tag{4.11}$$

To be able to use this function for estimation, CNLS uses axioms describing properties of the function f. In a simple example and for an analysis with a single input and single output, the function f is assumed to be monotonically increasing in x_i and concave. This leads to the following formulation of the estimation problem. The expression ϕ_i is used to denote the estimated value of $f(x_i)$ for firm i. The observations are ordered according to the input levels (from smallest to largest) so that the monotonicity and concavity could be formulated as in equation (4.12).

$$\begin{aligned}
\min_{\phi} \quad & \sum_{i=1}^{n}(y_i - \phi_i)^2 \\
\text{subject to} \quad & \phi_i \leq \phi_{i+1} \quad \forall i, ..., I-1 \\
& \frac{\phi_{i+1} - \phi_i}{x_{i+1} - x_i} \geq \frac{\phi_{i+2} - \phi_i}{x_{i+2} - x_i} \quad \forall i, ..., I-2
\end{aligned} \tag{4.12}$$

To perform the frontier estimation, the CNLS method can be adjusted so that all observations are located below or on the efficiency frontier. This adjustment, which leads to a method called corrected CNLS, is performed analogically to corrected OLS. That means that the residual of the most efficient firm is used to shift the constant and adjust the residual of all observations. However similarly to COLS, it may suffer from large sensitivity to outliers.

The efficiency analysis method based on CNLS is the StoNED method named in the introduction of this section. This method is performed in several steps. First the CNLS is estimated and residuals are estimated. To obtain the average and individual efficiency, either parametric methods (method of moments or quasi-likelihood estimation) or non-parametric kernel deconvolution[11] can be used. For the derivation of the individual scores the same procedure as for SFA (proposed by Jondrow *et al.* 1982) may be used.

As previously described for SFA, exogenous variables can be included in the StoNED analysis to account for the factors influencing inefficiency. The procedure is similar. Furthermore, the StoNED method can be adjusted for the use of panel data. Yet the problem becomes more complex. To lower the complexity further assumptions are made, such as no technological change, time-invariant inefficiency, and no correlation of the inefficiency with inputs.

The idea of combining the non-parametric feature with the stochastic framework brings new and promising aspects to the efficiency analysis. However, it does not seem to be fully equipped for complex efficiency analyses yet, especially when panel data are investigated. Therefore this method is not used in the current analysis.

[11] The kernel deconvolution is a method to estimate density of residuals based on the truncated distribution of inefficiency.

4.4 Comparison of the methods

In the chapter 3 and 4 we have presented a detailed summary of the methods used in efficiency analyses. With the exception of the last method, StoNED, these form two groups: the deterministic non-parametric methods (DEA, FDH, and PFA) and the parametric methods, which are mainly represented by the SFA method. Both groups have distinct advantages and disadvantages.

The SFA method is praised for including the stochastic error in the production equation. That allows the distance to frontier to be caused not only by inefficiency but also by stochastic noise. It can be used to study production as well as cost frontier. Moreover, there is a wide range of possibilities to study the panel data. On the other hand it is often criticised due to its assumptions on the production function and on the distributions of the errors. Assumptions may be too restrictive and may produce inconsistent results in case of incorrect use.

The non-parametric deterministic methods, DEA and FDH, do not need many assumptions apart from a few axioms on the production function and the assumption that the entire distance from the frontier is caused by inefficiency. While the axioms, such as strong disposibility, constant or variable returns to scale, and in case of DEA convexity, are not very restrictive, the last assumption is the base of the criticism of the DEA method. For it may not seem realistic to assume that there are no stochastic shocks and only deterministic inefficiency accounts for the deviations from the frontier.

DEA also used to be criticised for no possibility of stochastic inference because the linear programming techniques do not produce standard errors or confidence intervals for the efficiency estimates. However, the application of bootstrapping and the theory on asymptotic distributions solves this issue. Further problem of DEA and FDH, sensitivity to outliers, is solved by the PFA which is less influenced by extreme values because it allows a small portion of observations to lie above the frontier. DEA uses window analysis and Malmquist index to cope with panel data. While window analysis may be sometimes restrictive, DEA analysis based on Malmquist index allows both technological change and individual deviations from the frontier to vary over years.

To sum up, both DEA and SFA have their advantages and disadvantages. For efficiency analysis in general and for the public sector efficiency analysis in particular, none of the methods seems to be clearly preferable. DEA produces results which do not depend on functional and distributional assumptions. FDH applies even less assumptions and may be used the confirm clear inefficiency among the observations. On the other hand, SFA builds on the statistical and probabilistic framework to model the production process. When inputs are in the form of expenditure, SFA estimation of cost frontier may be applied to utilize this form of data. Cost frontier also seems to be convenient for the public sector analysis if particular inputs are not clearly defined. Furthermore, for deeper study of panel data, SFA offers more complex techniques. Thus while generally the parametric and non-parametric methods may be viewed as complements, for the current and similar analyses SFA using cost frontier approach seems to be the most suitable alternative available.

Therefore the main method used in this study is the cost frontier approach of SFA. Several panel data models are used to produce the results. These are specified in the chapter 6, which presents the results. Nevertheless, applying more methods gives a richer view on the analysed sample. Therefore the main results are compared to a variety of available methods for efficiency analysis including both the deterministic methods as well as different specifications for the SFA analysis.

Chapter 5

Data description and regional services

This chapter presents the data used to study the general spending efficiency of Czech administrative regions (in Czech “kraje”). The first part of the chapter briefly describes the organization and services provided by Czech regional offices. The second section introduces the single input variable, which is expenditure as in most of the previous studies. In the next section output side is discussed. As the output variable we create an index which aims to capture the provision of services by regions. Within the section, individual indicators as well as the way to construct the index from these indicators are described. Last section covers several variables that are studied for the influence on efficiency scores.

5.1 Czech regions and their services

There are 13 regions in the Czech Republic and one capital city with regional status, the capital city Prague. This division corresponds to level 3 in the system of Nomenclature of Units for Territorial Statistics (NUTS). The self-governing regions in the current form were established after elections in 2000. In 2003, the regions assumed the responsibilities of the previously used smaller districts (in Czech “okresy”).

The main executive body on the regional level is regional board (“rada kraje”). The main decision-making body on the regional level is regional council (“zastupitelstvo kraje”), elected every four years. The main authority of the regional council and board is provision of regional government services. The council and the board are assisted by regional office ("krajský úřad"), which also provides delegated central government services.

Czech regions are financed by subsidies from the Czech government budget, along with tax revenues, European grants, their own revenues, as well as loans and credits when needed. Significant part of expenditure covers the provision of public goods and services. The services provided by the regional offices can be divided into several areas: transportation, regional development, spatial planning, social services, culture, education, health, environment, general administration and other services. To specify the services, we name some of the most important agenda of the regions.

In the area of transportion, regions maintain and repair roads in their territory, and coordinate and facilitate the availability of transport services. The department of regional development focuses on preparation and cooperation on European projects, supporting among others rural landscapes, small and medium enterprises, and tourism. In terms of spatial planning, regional offices supervise municipalities in their authority to issue building permits and publish Principles of territorial development ("Zásady územního rozvoje"). Regions also provided social services such as child protection, financing of retirement and nursing homes, nursing services, and other services for economically or socially disadvantaged people. Furthermore, regions finance museums, galleries and other cultural projects, along with the conservation and preservation of cultural heritage.

One of the areas with significant expenditure is education because regions organize secondary education.[12] Apart from the organization of secondary education, the regional education department prepares and publishes long-term plans for education and development of education system in the region, and supports other youth and sports activities and education of the public.

In the area of health, regions primarily establish emergency medical services as semi-budgetary organizations. Besides, regions provide supervision and registration of medical services. The department of the environment is in charge of environmental subsidies and regulation of noise, emissions and other sources of pollution and contamination. Further duties include ecological education, maintenance of protected areas (excluding national parks), and registration of rare species.

General administration includes, among others, registration of births, deaths

[12]In ISCED (International Standard of Classical Education) Czech secondary education corresponds to higher secondary education, ISCED 3.

and migration of residents. As other services, let us name, for example, cooperation with employment offices and realization of European grant projects in the area of human resources. Furthermore, regions coordinate international interaction and cooperation. Regional offices also include an information technology (IT) department, which takes care of internet communication and safety, information and data management, and maintenance of hardware and software of the regional offices.

5.2 Input variable - expenditure

The analysis of efficiency of public sector, as compared to the private sector, offers several issues. In contrary to the private firms, for the public sector it may not be entirely clear what input enter the "production process" and, even more importantly, what are the outputs. Yet, to obtain reasonable results inputs and outputs need to be defined and measured properly.

Typically, the accounts of public organizations do not inform us about resources used for particular purposes. Some authors (such as Worthington 2000) use apart from financial expenditure also the number of civil servants as a labour input and non-financial expenses as a capital input. These data are, however, typically not available and thus a single monetary measure is used instead to account for the cost of services. This is the case of the current analysis as well. In particular, we use the consolidated current expenditure of regional offices (in Czech: "konsolidované běžné výdaje krajského úřadu"). The reason for the choice of current expenditure lies in the relevance to the current year. Real expenditure, expressed in the prices of 2005, is used to enable the comparison across years.

Data are available for all 14 Czech regions since 2001 when the regions started to operate as individual administrative units. However only data from years since 2003 are used because during the first two years both duties and expenditure of regions were phased in. By the end of the year 2002 districts ceased to exist as self-governing units and their duties were given to regions. The databases of the Ministry of Finance of the Czech Republic are the source of the data. The table 5.1 displays the summary statistics of the expenditure for all regions separately.

Table 5.1: Summary of real expenditure across regions

code	region name	mean	st. deviation	minimum	maximum
CZ010	Hl. m. Praha	37432977	2065393	35282564	42790788
CZ020	Středočeský kraj	12879549	1051420	10821354	14751301
CZ031	Jihočeský kraj	7713165	551759.5	6785337	8832817
CZ032	Plzeňský kraj	6155366	461874	5261442	6697527
CZ041	Karlovarský kraj	3577816	197499.2	3180006	3784255
CZ042	Ústecký kraj	9071885	602998.1	7926744	9874679
CZ051	Liberecký kraj	4945444	294591.8	4361607	5254847
CZ052	Královéhradecký kraj	6496776	407110.7	5668529	7000893
CZ053	Pardubický kraj	5739321	302711.1	5027581	6231585
CZ063	Vysočina	6322632	468071.6	5539597	6869109
CZ064	Jihomoravský kraj	11484568	584704.8	10130321	12062725
CZ071	Olomoucký kraj	7274827	401470.8	6465807	7823213
CZ072	Zlínský kraj	6305786	225998.8	5984675	6784698
CZ080	Moravskoslezský kraj	12340595	565653.7	11265207	13175292
CZ	Czech Republic	9838622	8165727	3180006	42790788

Source: Author's computations based on data from the Czech Ministry of Finance. *Note*: in thousands of 2005 CZK, means 2003–2014.

5.3 Output indicators

As noted in the previous section, even more problematic may be the search for appropriate output variables. Similar to De Borger & Kerstens (1996), Vanden Eeckhaut *et al.* (1993), Kalb (2010), or Worthington (2000) this study attempts to analyse the efficiency of general public spending in order to obtain the overall picture, and does not just focus on one particular area, such as education or health care. Therefore, the output side should reflect various services provided by the regional office. Depending on the data availability, previous studies select several indicators that may serve as a proxy for the relevant public services.

There is a slight difference between the previous studies and the current one. The previous works that study administrative units smaller than whole countries concentrate usually on municipalities. On the other hand, the focus of the current analysis is on the regions, administrative units larger than municipalities but smaller than states. The services of regions, which have been summarized in the first section of this chapter, are different than of municipalities. Therefore the indicators, though inspired by previous works, are mostly different as well. The list of all output indicators along with an explanation for each follows directly.

- **Total number of residents** - proxy for services distributed randomly along the population which could not be accounted for by any other indicator, such as general administration and education of the public.

- **Total area** - proxy for services distributed over the whole area which could not be accounted for by any other indicator, such as transport availability, support for rural landscapes, publishing of Principles of territorial development, and protection of the nature.

- **Total number of personal automobiles** - proxy for maintenance and repairing of roads.[13]

- **Share of registered entrepreneurs to total number of registered economic subjects** - higher share of individual entrepreneurs indicates better opportunities for smaller firms; a proxy for support for small and medium businesses.

- **Number of guests in collective accommodation establishments without residents** - indicator of the amount of tourists visiting the region; it measures tourism support, cultural projects organization, and cultural heritage protection.

- **Total number of issued building permits** - proxy for supervision of building departments of municipalities.

- **Number of residents aged 65 years and above** - proxy for the demand for retirement and nursing homes and nursing services for older people.

- **Number of residents 14 years and below** - proxy for child protection services as well as for supporting organizations for youth and sport.

- **Students in secondary schools as the share of the population in the age of 15-18 years old** - indicator of the extent to which secondary schools are attended in the region.

[13] Even better indicator for this services would be total length of roads maintained by the region. However, this indicator was available only since 2007 which would significantly decreased the length of data and thus cause problems for the analysis due to low number of observations.

- **Number of doctors in the emergency medical services** - indicator of the range of the service of emergency medical services provided.[14]
- **Emissions per square kilometre (particulate matter[15])** - indicator of the extent of air pollution showing the need for environmental care and pollution regulation.
- **Immigrants per 1000 residents** - with higher rate of newcomers in the region, there is more administration (in terms of registration).
- **Unemployment rate** - indicator of the need for promotion of employment in terms of European human resources projects and cooperation with employment offices.
- **Share of households equipped with personal computer** - with higher computer literacy in the population, there are more services provided on internet and generally in electronic form which causes higher need for all services of the IT department.

However, it must be noted that these indicators are mostly only proxies for the public services. This is partly caused by the nature of inputs and outputs in analyses of public services. This problem could be mitigated if more appropriate data were available, measuring the exact amount of the services or even better the quality of the services. Yet, to our knowledge such data are not available. Therefore we cannot account for the differences in quality as they are unobserved.

The output indicators are transformed to an index such that the non-parametric methods (especially FDH) do not lose the power to reject efficiency. When there are too many outputs, the non-parametric methods compare along too many dimensions and thus only few observations may be dominated in terms of efficiency. The parametric methods lose degrees of freedom in case of large number of variables. The index is formed similar to Afonso & Fernandes (2006). Firstly, the output indicators are proportionally transformed to have mean equal to 1 so that they are comparable to each other. Then they are summed up with equal weights as there are no intuitive weights which could be assigned.

[14]Again better indicator would measure the performance of the service, such as at least the number of patient treated. Yet this indicator was available for only a few years.

[15]There are data on four different emissions available: particulate matter, sulphur dioxide, carbon oxide and oxides of natrium. However they are highly correlated. Particulate matter was chosen because it has highest correlation with other measures.

Most of the output indicators used are from the Czech Statistical Office where they are published in the form of regional time series for years 1993–2014. Missing observations were added from regional yearbooks and the Public Database of the Czech Statistical Office when available. Current data on education were added from the Ministry of Education, Youth and Sports. Lastly, data on emergency doctors are from the Institute of Health Information and Statistics of the Czech Republic.

Some missing data could not be found and were projected using the information available. We could not find any data on the share of households with personal computer in 2003. Therefore the data were fitted using logistic function of time trend and dummies for each region (with R-squared = 0.9730) and the data for 2003 were predicted using the estimated values. Similarly, data could not be obtained for emissions in 2014 and were predicted using linear trend for last years and regional dummies[16] with even higher variation explained (R-squared of the linear regression is 0.9979).

5.4 Variables explaining inefficiency

The estimated inefficiencies can have different causes. When a unit is estimated to be inefficient, the reason need not lie only in suboptimal management. The process may be affected by exogenous input variables, as well as unobserved output variables. Variables described in this section will be used to investigate several hypothesis on the reasons behind the estimated inefficiency. These reasons are connected both to the managerial performance and to the environmental effects, thus in some cases it might be difficult to decide whether a variable serves as effect on sub-optimality of management or as an exogenous input or output.

First, in regions with higher average age, it may be more costly to provide public services. Also older people tend to prefer traditional ways to optimization. Therefore average age is expected to negatively influence efficiency. Second, it may be more difficult to provide optimal services in regions with lower density of population because it may be more costly to reach all clients of regional services when they live far from each other. Thus density could affect efficiency positively.

Third, it could be argued that poorer regions, which have more tight budget,

[16]Specific time trend for Prague and Moravian-Salesian region had to be allowed because both had very high values at the beginning of the examined period.

concentrate more on optimization. This is proxied by gross domestic product (GDP) per capita which could thus be negatively correlated with efficiency. It might also be expected that on average poorer people have higher incentives to monitor public expenditure. On the other hand, this hypothesis could be opposed by an argument that in regions with higher GDP per capita some residents have better financial education and thus can monitor the use of public expenditure. This would indicate a positive effect. Moreover, some might claim that regions may be richer when they spend money more carefully. This would indicate even more complicated causality. The effect of richer and more financially educated people could be measured by average monthly gross wage but there is high correlation between regional GDP per capita and average monthly gross wage. Thus only one indicator, the GDP per capita, has been chosen.

Fourth, the monitoring of public expenditure may be correlated also with general education of the residents which could potentially account for part of the ambiguous effect of the wealth (indicated by GDP per capita). More educated population has more incentives and higher abilities to assess and criticise the activities of public organizations. Thus the share of population with lower secondary or lower education ("ukončené základní vzdělání") is expected to be negatively correlated with efficiency.

Last, efficiency may reflect the political situation in the region. The effect of the share of selected political parties on the efficiency of spending can be studied. For this analysis the parties selected are the Communist Party of Bohemia and Moravia and Czech Social Democratic Party, which represent the main parties in the left wing of the Czech political spectre.[17] This effect is indicated by the separate share of each of the two parties in the coalition at the regional council. Separate shares are used instead of the sum of shares for both socially oriented parties because it is a priori not clear whether these parties are more complements or competition. It may be assumed that the social democrats and especially communists tend to spend larger amounts of expenditure in comparison to other parties. The typical voters of the communist and social democratic parties in the Czech Republic may be older people with lower education, but these effects are already explained by the average

[17] The reason for the choice lies in the fact that these two parties belong to the strongest Czech political parties in the regional councils throughout the whole analysed period of time. The only remaining party with similar power and persistence is the Civic Democratic Party but it was excluded due to very high negative correlation with the fractionalization index defined below.

age of population and percentage of population with lower education. Thus the effect is not clear.

Different hypotheses concern the shares of parties in the regional council. It may be argued that higher number of political parties in the regional council can lead to inefficiency due to more difficult communication, longer decision making and more compromises. On the other hand, higher number of parties with lower shares may lead to decreased corruption and increased monetary benefits due to higher mutual control and more difficult collusion. A positive effect of additional political challenger in Czech municipal governments on efficiency is recently confirmed by Palguta (2015).

The measure chosen is the fractionalization index, which is suggested by Rae (1968). Fractionalization may be defined as the probability that randomly chosen pair of politicians belong to different political parties. The index is defined as $F = 1 - \sum_{i=1}^{m} t_i^2$ where t_i denotes the proportion of council members associated with the party i.

For most of these variables, Prague has significantly different values than the rest of the Czech Republic. In Prague there is slightly lower percentage of people with low education and slightly higher average age (especially in the first years). There is significantly higher GDP per capita. Not surprisingly, the highest difference is for the density of population as, in contrary to the rest of regions, Prague is a city. The mean density without Prague is twenty times smaller (122.8 residents per kilometre squared) than the mean score for Prague (2459.8). When Prague is excluded, the Moravian-Silasian region stands out with the density of 228.1 people per square kilometre as compared to the mean of the remaining 12 regions which is 112.1. Thus to determine the effect of density, the results have to be compared for the regions without Prague and Moravian-Silesian region. For the rest of the variables, the effect is checked against the results from analyses without Prague as well.

Data on these explanatory variables come mostly from the regional time series from Czech Statistical Office. Information about low educated people is taken from Public Database of Czech Statistical Office. Data on shares of political parties in the regional council come from the official database on election results, www.volby.cz. The election to the regional council takes place every four years since 2000. The data on corresponding elections in the capital city Prague, elections to the council of the capital city Prague, were taken for years 2002, 2006, 2010 and 2014.

Chapter 6

Analysis

This chapter contains the discussion of the results of the efficiency analysis. First, the estimated efficiency scores are presented and interpreted. Then follows the summary of alternative results to study the robustness of the results. The chapter is concluded with a discussion of the effects of selected explanatory variables on the efficiency scores for the main results as well as for the alternative ones.

6.1 Main results

The previous two chapters examined in detail the methods for efficiency analysis. Based on that discussion and the form of data, Stochastic Frontier Analysis method using cost frontier approach for panel data is chosen as the main tool. The scores are estimated adopting the input orientation because while regions may adjust the costs, their ability to influence the output index is not clear. Thus the estimated efficiency scores capture the input efficiency.

The main method varies with the assumption on the time variation of efficiency. If the efficiency scores are assumed to be constant over time, fixed effects or random effects approaches to panel data may be applied. For the estimation of time-varying efficiency, more complex models are used. The approach suggested by Cornwell *et al.* (1990), where time-evolution of efficiency is estimated for each DMU separately, is not selected due to relatively low number of observations in comparison to the high number of explanatory variables needed. The next most flexible models are those that do not assume any functional behaviour of the time-varying efficiency and can separate it from the persistent efficiency. The model introduced by Kumbhakar & Heshmati (1995) (hereafter KH95) assumes that a part of inefficiency is permanent and a part varies over time. Thus it can separately estimate the constant and time-varying component of efficiency score. The model by Kumbhakar *et al.* (2014) (hereafter KLH14) even separates these two forms of inefficiency from the individual heterogeneity effects that are typically assumed in panel data models.

6.1.1 Time-invariant inefficiency

The model for the cost frontier analysis, equation (6.1), is a panel data extension of equation (4.9), where the cost frontier is assumed to be a translog function. This specification satisfies the required monotonicity in output of the cost frontier (as is introduced in 4.2.2) for all observations.

$$ln(C_{it}) = \beta_0 + \beta_y ln(y_{it}) + \frac{1}{2}\beta_{yy}(ln(y_{it}))^2 + v_{it} + \eta_i, \qquad \text{(FE/RE)} \qquad (6.1)$$

In case of the assumption of time-invariant inefficiency, the frontier and the efficiency scores are estimated using classical panel data approaches, fixed effects and random effects. These two models differ in the assumption on the independence between the explanatory variables and the individual heterogeneity, which in this case coincides with the inefficiency. It is not clear from the theoretical point of view whether this independence holds, that is whether inefficiency is not correlated with the amount of output in first and second power. When the inefficiency is independent from the explanatory variables, both methods are consistent and random effects are efficient. However when the independence does not hold, only fixed effects are consistent and thus possible to use. This hypothesis can be tested using the Hausman test. For the specification used in this thesis, the fixed effects approach is preferred based on the results from the Hausman test.

The fixed effects specification allows the explanatory variables to be correlated with efficiency scores without causing bias to the estimates. However, as only a small number of regressors are used, it may happen that the explanatory variables are correlated with an omitted variable included in the random error term. To test and correct for this alternative the panel data approach to instrumental variables may be applied. For the instrumental variables estimation a relevant instrument uncorrelated with the explained variable is necessary.

As the instrument for the main analysis we have chosen the number of pupils in primary schools. This is correlated with the output index which includes the number of residents, number of children, as well as the share of secondary school students to the corresponding population. On the other hand it is not correlated with the regional spending because primary schools are financed by municipalities. Both these correlations were tested using panel data regressions. For the squared logarithm of output index squared number of primary school pupils is selected.

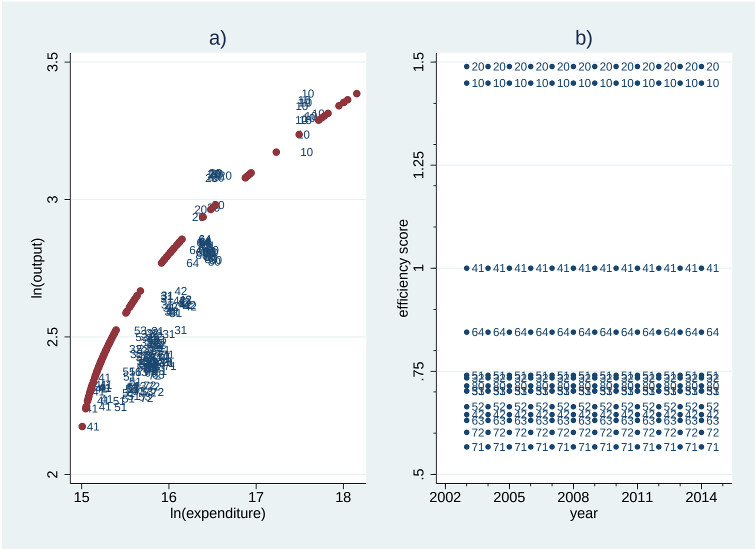
Applying the instrumental variables regression, the regressors, logarithmic output in first and second power, were tested for endogeneity. The hypothesis of exogeneity was rejected even on 1 % significance level. The regression is exactly identified, using two instruments for two endogenous variables. None of other potential instruments available, including doctors in hospitals or crime rate (both satisfying the necessary conditions for a suitable instrument), is found an appropriate improvement using the Sargan statistic for over-identification test of instruments. Using the Anderson (1951) canonical correlations test, the instruments are tested to be relevant. The instruments are also checked for the weak identification using Stock-Yogo weak ID test (Stock & Yogo 2005). The resulting statistic compared to the reported critical values shows that the instruments are not only weakly identifying.

The results are presented in the form of two graphs. Figure 6.1 a) depicts the efficiency frontier. The figure 6.1 b) shows the efficiency scores of the regions for all years.[18] The efficiency frontier is estimated without the observations for Prague to eliminate the possibility that it causes bias to the frontier. Nevertheless the inclusion of the capital city changes the frontier only slightly and Prague is still estimated to be efficient.

The average efficiency score is 0.820. The minimal score (0.567) belongs to the Olomouc region and the maximal score (1.489) is reached by the Central Bohemian region, closely followed by the capital city of Prague (with a score of 1.449). The frontier is shifted to exclude outliers that would otherwise shift the efficiency scores significantly downward. The outliers, in this case, are the two most efficient regions, Prague and the Central Bohemia. Inclusion of these regions below the frontier would decrease the average efficiency score by 27 percentage points (to 0.551). This approach, similarly to PFA, aims to ensure lower sensitivity to extreme points which leads to higher robustness of the results. As can be seen later, in comparison with the further main results

[18]This form of presentation of results is chosen to provide better comparison with further resulting efficiency scores.

Figure 6.1: Time-invariant inefficiency, frontier and $exp(-\hat{\eta}_i)$ from (6.1)



Source: Author's computations. *Note:* The regions are denoted in the following way: 10 = capital city Prague, 20 = Central Bohemian, 31 = South Bohemian, 32 = Plzeň, 41 = Karlovy Vary, 42 = Ústí nad Labem, 51 = Liberec, 52 = Hradec Králové, 53 = Pardubice, 63 = Vysočina, 64 = South Moravian, 71 = Olomouc, 72 = Zlín, 80 = Moravian-Silesian.

(section 6.1.3 and figure 6.4), the high dispersion of the efficiency scores may be caused by heterogeneity of the regions.

The most efficient region, the Central Bohemian region is one of the largest regions with the highest number of residents. The region has relatively small expenditure in comparison to its size which places it close to the frontier. It may be related to the fact that it surrounds the borders of the capital city of Prague and thus some of its services, for example the secondary education, are provided by Prague.

The capital city of Prague seems to be an outlier. Even though it lies above the frontier, it has very high costs in comparison to the rest of the regions. Along with the notion of Prague providing services not only for its residents, the reason for this observation may lie in the fact that Prague has a specific position among the regions. Being the capital city, it serves as a region and a municipality at the same time. Thus it finances not only the services provided typically by regions, but also the municipal services. Nevertheless even with the exclusion of the Prague from the estimation of the frontier, the efficiency estimation assigns it the second highest efficiency score. It could be hypothesized that Prague as a capital city with the highest expenditure is closely observed by citizens as well as politicians. It is relatively difficult to compare Prague to the rest of the region as it has a different position. Yet we still include it for completeness and comparison.

6.1.2 Time-varying inefficiency: KH95 model

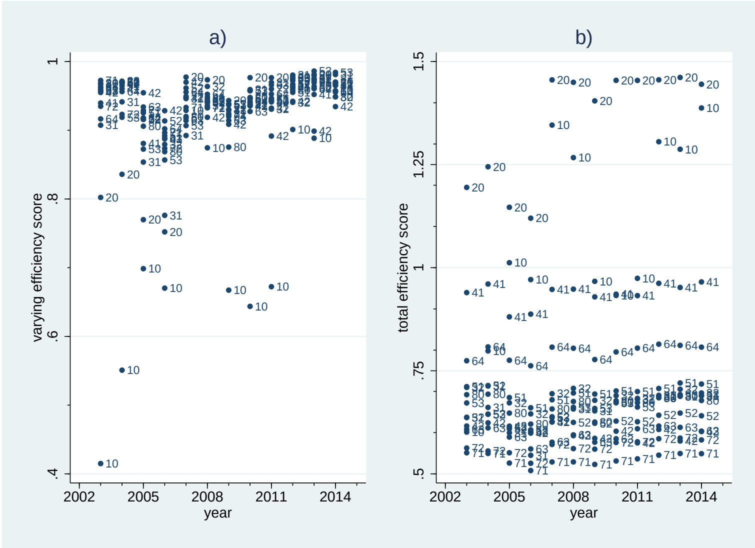
The following equation (6.2) shows the KH95 model which allows the separation of persistent (η_i) and time-varying (τ_{it}) inefficiency. From the previous model it differs by extracting the time-varying inefficiency from the composite error. Thus the persistent inefficiency is identical to the time-invariant inefficiency estimated previously. Therefore we do not present the persistent efficiency scores as they are already presented in the figure 6.1 b). The time-varying efficiency is estimated using MLE estimation along with distributional assumptions on the non-constant error terms.[19] Total inefficiency may be expressed as the product of the persistent and time-varying technical efficiency (TE): $TE_{TOTAL} = TE_{VARYING} \times TE_{PERSISTENT} = exp\{-(\eta_i + \tau_{it})\}$. The results for time-varying and total efficiency are presented in figures 6.2 a) and b).

$$ln(C_{it}) = \beta_0 + \beta_y ln(y_{it}) + \frac{1}{2}\beta_{yy}(ln(y_{it}))^2 + v_{it} + (\eta_i + \tau_{it}), \qquad \text{(KH95)} \quad (6.2)$$

To compare the regions based on the results it can be noted that with the exception of Prague the order of the regions based on the total efficiency score is not different from the order already discussed in the previous subsection, for the time-invariant efficiency scores. The reason lies in the fact that for all regions, with the exception of Prague and Central Bohemia, there is relatively little time-varying inefficiency and it is very similar between the regions. The average time-varying efficiency is 0.923, which leads to the average of the total efficiency score equal to 0.747. The time-varying efficiency is significantly different from one (on the significance level $\alpha = 0.05$) although for most of the regions only marginally. The scores have relatively small dispersion with a few jumps which occur, apart from Prague and Central Bohemia, for a few observations from the South Bohemian and Ústí nad Labem regions. The time-varying efficiency score is on average lower in the years 2005, 2006, and 2009 even when Prague

[19] The random term, v_{it}, is assumed to be normally distributed ($v_{it} \sim N(0, \sigma_v)$) and the efficiency term, τ_{it} is assumed to follow half-normal distribution ($\tau_{it} \sim N^+(0, \sigma_\tau)$). Truncated normal distribution was tested for the τ_{it} term but the mean was found not significantly different from zero.

Figure 6.2: KH95 model, $exp(-\tau_{it})$ and $exp\{-(\eta_i + \tau_{it})\}$ from (6.2)



Source: Author's computations. *Note:* The regions are denoted in the following way: 10 = capital city Prague, 20 = Central Bohemian, 31 = South Bohemian, 32 = Plzeň, 41 = Karlovy Vary, 42 = Ústí nad Labem, 51 = Liberec, 52 = Hradec Králové, 53 = Pardubice, 63 = Vysočina, 64 = South Moravian, 71 = Olomouc, 72 = Zlín, 80 = Moravian-Silesian.

is disregarded. This occurrence coincides with a slight overall increase in real expenditure in the same years, while output remained stable.

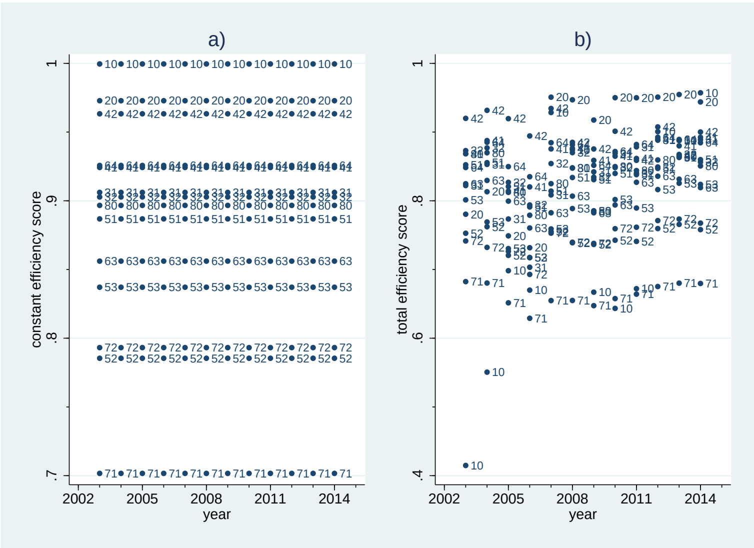
For Prague, the time-varying efficiency (and hence also the total efficiency) increases significantly over the studied time period from 0.4, staying longer around the score 0.7 and at the end reaching a time-varying score close to 1. This is caused by the fact that Prague's real expenditure changed relatively little while the output score increased significantly. While the total number of residents and number of children increase in Prague slightly more than in other regions, the increase of output index is mainly driven by the significant increase of the share of secondary students and the amount of tourists visiting the capital city.

6.1.3 Time-varying inefficiency: KLH14 model

The previous model, KH95, confuses individual regional effects with the persistent inefficiency. Consequently the efficiency scores may have a downward bias because the whole time-invariant part is assumed to be inefficiency. This potential bias is corrected in the last model which contains four components of the error. Not only can this model separate persistent (η_i) and time-varying (τ_{it}) inefficiency. It can also isolate these inefficiencies from the heterogeneity of individual units (a_i). The last term, v_{it}, is the random shock. To estimate all three efficiency scores - persistent, time-varying, and total score - two MLE estimations are used. The scores are identified based on distributional assumptions on all four components of the error term.[20]

$$ln(C_{it}) = \beta_0 + \beta_y ln(y_{it}) + \frac{1}{2}\beta_{yy}(ln(y_{it}))^2 + a_i + v_{it} + (\eta_i + \tau_{it}), \qquad \text{(KLH14)} \tag{6.3}$$

Figure 6.3: KLH14 model, $exp(-\eta_i)$ and $exp\{-(\eta_i + \tau_{it})\}$ from (6.3)



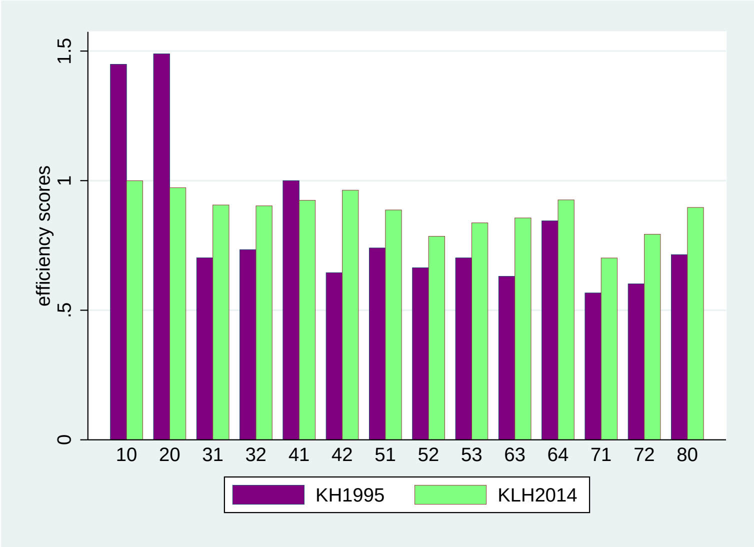
Source: Author's computations. *Note:* The regions are denoted in the following way: 10 = capital city Prague, 20 = Central Bohemian, 31 = South Bohemian, 32 = Plzeň, 41 = Karlovy Vary, 42 = Ústí nad Labem, 51 = Liberec, 52 = Hradec Králové, 53 = Pardubice, 63 = Vysočina, 64 = South Moravian, 71 = Olomouc, 72 = Zlín, 80 = Moravian-Silesian.

The KLH14 model differs from the previous one, KH95, by the separation of the heterogeneity and persistent inefficiency. Therefore the time-varying efficiency score is not presented again because it is identical as for KH95. The figure 6.3 shows the results for the persistent component (which is newly separated from the heterogeneity) and the total efficiency scores.

From the figure 6.4 it can be seen that the persistent efficiency changed. The persistent efficiency scores for the KH95 model (which are identical to the time-invariant efficiency obtained in FE) are much largely dispersed with the scores for Prague and Central Bohemia located far above the usual efficiency

[20]In addition to the previous assumptions, a_i is assumed to follow normal distribution ($a_i \sim N(0, \sigma_a)$) and η_i is assumed to follow half-normal distribution ($\eta_i \sim N^+(0, \sigma_\eta)$). The hypothesis of truncated normally distributed η_i is again rejected based on insignificant mean.

Figure 6.4: Results for the KLH14 model - time-invariant score, comparison to KH95 $\sim$ FE



Source: Author's computations. *Note:* The regions are denoted in the following way: 10 = capital city Prague, 20 = Central Bohemian, 31 = South Bohemian, 32 = Plzeň, 41 = Karlovy Vary, 42 = Ústí nad Labem, 51 = Liberec, 52 = Hradec Králové, 53 = Pardubice, 63 = Vysočina, 64 = South Moravian, 71 = Olomouc, 72 = Zlín, 80 = Moravian-Silesian.

limit of 1 and the lowest scores lying below 0.6. On the other hand, for the persistent component separated from the individual heterogeneity the efficiency scores do not need to be shifted to account for outliers. Thus it appears that the high dispersion of the KH95 efficiency scores is caused by the heterogeneity of the regions, not accounted for in the KH95 model. For the KLH14 the maximal score (1.000) is reached by Prague, closely followed by the Central Bohemian region (0.973). Thus the first two positions are occupied by the same regions only in the reversed order. Similarly, the last two positions are left for the Olomouc and Zlín regions. The average score is 0.882 which is larger even than the previous average persistent efficiency score which was obtained after shifting the frontier. Thus it seems that the KLH14 model is much less affected by outliers.

The order of the regions based on the persistent efficiency score is again more or less maintained for the total efficiency scores as the time-varying efficiency (same as for KH95) has relatively low variance and similar movement for the regions. The exceptions are the two most efficient regions, Central Bohemia and the capital city of Prague which reach their prime positions in the later periods (in case of Prague even in the last period). Thus in the first periods the most efficient seems to be the Ústí nad Labem region. This might be a surprising finding due to the fact that for the KH95 model the region of Ústí nad Labem is one of the least efficient regions. It implies that the region of Ústí nad Labem has some adverse regional characteristic which can be separated from inefficiency as a (randomly distributed) regional effect.

6.1.4 Discussion of the main results

To summarize the main results, the efficiency scores are varying over time, as the time-varying component is significantly different from one. Nevertheless the time-variation is relatively small for most of the regions. The highest change occurs for Prague and Central Bohemia which both significantly improve their performance over the studied time period. Furthermore, it can be noted that there is some estimated heterogeneity which causes the differences between the persistent scores for KH95 and KLH14 models. For the three most efficient regions according to the KH95 estimation, the KH95 score is higher than the KLH score. However this is caused by the frontier shift used in the KH95 estimation to account for the extreme values. Otherwise the score would lie below the score estimated by the KLH14 estimation, as is true for most of the regions. This confirms the downward bias of the KH95 persistent efficiency scores caused by the inclusion of the heterogeneity.

Based on the regional comparison, the Central Bohemian region is estimated to be the most efficient region with the exception of the year 2014 when it is surpassed by the capital city of Prague. While Prague holds one of the prime positions, it has the largest time-variation of efficiency scores due to the large increase in the output index score. Furthermore, to the more efficient regions belong the South Moravian region, along with the Karlovy Vary region. The separation of the heterogeneity from the persistent efficiency scores favours some of the regions significantly more than the KH95 model. An example is the Ústí nad Labem region which changes its position among the regions from the fourth least efficient to the third most efficient. This may indicate that the region performs efficiently but suffers from adverse environment.

The regions may be ranked from the most efficient to the least efficient using the efficiency scores. However the application of the results from efficiency analysis focuses rather on the investigation of the reasons and conditions leading to inefficiency. The regions that are estimated to be more efficient are not situated at one part of the Czech republic. Nevertheless, some of them lie close to each other. As was noted before, the region of Central Bohemia surrounds the borders of the capital city of Prague, affecting some of the indicator scores for these two regions. Additionally, Karlovy Vary and Ústí nad Labem regions are both in the north-western part of Bohemia, on the border to Germany and thus they may have similar environments. Similarly, among the least efficient regions are consistently the Olomouc and Zlín regions which are neighbours and thus might be influenced by similar problems. The regions may also have some historical or social problems which leads to inferior position. In this thesis we do not aim to find all the causes of inefficiency. Only selected variables are studied. The results are presented in the last section of this chapter (section 6.3) after the discussion of results from alternative estimators, which are presented in the following section.

6.2 Robustness study

In this section results using various different methods for efficiency estimation are considered. Firstly the non-parametric methods, DEA, FDH, and PFA, are analysed. Then the results are compared to SFA analysis using the production frontier approach, instead of the cost frontier. Lastly, cost frontier approach is presented for pooled data as there is relatively small variance in the panel data.

Table 6.1: Results from alternative methods

method	mean score	st. error	minimum	maximum	# of efficient
DEA	0.770	0.127	0.533	1.025	15
			Olomouc 06	Prague 08	(3 above)
FDH	0.919	0.088	0.598	1	39
			Zlín 08	(39 observations)	
PFA	0.965	0.116	0.701	1.594	61
			Zlín 06	Karl. Vary 07	(23 above)
SFA prod.	0.640	0.152	0.400	1.159	9
pool COLS			Olomouc 06	(Centr. Bohemia)	(8 above - 5%)
SFA prod.	0.865	0.191	0.427	1.341	24
panel FE			(Prague)	(Karl. Vary)	(12 above)
SFA cost	0.708	0.137	0.460	1.124	9
pool COLS			Prague 03	(Centr. Bohemia)	(8 above - 5%)
SFA cost	0.885	0.181	0.212	0.986	0
pool MLE			(Prague)	(Karl. Vary)	

Source: Author's calculations.

The results for the non-parametric methods are acquired using window analysis with three-year windows of data. Window analysis is selected to enable comparison between all three non-parametric methods. The non-parametric estimators are applied to the data in levels, not in logarithms, from definition. PFA is estimated using α-frontier approach with $\alpha = 95$ so that 95% of the observations lie below the estimated frontier. Only efficiency scores, not slacks, are reported for DEA since no significant slacks are present due to the low dimensionality.

The parametric methods are performed using the instrumental variable regressions, same as for the main results. For the production frontier estimation the instrument used is environmental expenditure in the region. The correlation with the real regional expenditure and the independence from the output index is tested. Results from all the tests are satisfactory. Where the maximal efficiency score lies above 1, the frontier was shifted to exclude the largest values (in case of pooled analysis - the largest 5 %, in case of panel data analysis - the region with the largest score). The results from all the alternative methods are presented in the table 6.1 and in the corresponding graphs in the Appendix A.

All the non-parametric methods have by construction high average efficiency scores and non-zero number of efficient units. While the DEA has smaller average than the main results, FDH and PFA have both average efficiency score higher than 90%. Thus those results serve rather as confirmations of inefficiency for the most inefficient regions. Similarly to the main results, Central Bohemia and Prague are found efficient by all the non-parametric methods. However, contrary to the main results, the region Karlovy Vary is estimated to be efficient as well.[21] A few observations from South Moravian, South Bohemian, Plzeň, and Liberec region have very high efficiency scores as well. Among the most inefficient according to the non-parametric methods belong the Hradec Králové, Vysočina, Olomouc, and Zlín regions.

The region of Karlovy Vary is one of the smallest regions in the Czech republic with the lowest number of residents. It has a very low score of the output index but also the lowest amount of costs. Thus it stands out from the other regions. It might potentially have more resources due to foreign financial help and tourists' investments. Nevertheless, the efficiency of the region could

[21] In each window of data for DEA some observation on the region of Karlovy Vary is even estimated to have efficiency score equal to one both under the assumption of variable returns to scale and the assumption of constant returns to scale. The remaining observations are estimated to operate under decreasing returns to scale.

also be due to the choice of method because the non-parametric methods tend to estimate slightly higher efficiency scores.

The alternative SFA methods present various results. The average efficiency score is relatively high, for some methods even slightly higher than for the main results. Based on all alternative estimates by SFA, Karlovy Vary region is among the most efficient regions. The region of Central Bohemia is among the most efficient with the exception of the MLE cost approach where its scores vary significantly. Prague is mostly seen as one of the least efficient. Based on the MLE cost estimation it is even extremely inefficient in comparison to the rest of the regions. Olomouc and Zlín regions are in almost every case among the least efficient regions.

To compare the main results to the alternative pooled data approaches for cost frontier (COLS and MLE), it can be noted from the graph A.6 that the estimated frontier using the first approach (COLS) is similar to the main findings. The difference lies in the slight rotation of the frontier which cause the capital city to be among the last on the scale of regions ordered according to the efficiency and its place as the second most efficient region occupied by the Karlovy Vary region. The results from the application of MLE estimation to the pooled data are similar to the findings for the time-varying efficiency for KH95 and KLH14. The efficiency is highest in the first and in the last years of the examined period. There are slight decreases accompanied with higher variance in the years 2005–2007 and partially in 2009 (for Prague in the years 2006 and 2010). The similarity of the MLE efficiency scores to the time-varying score from the main results may suggest a different approach of the MLE efficiency estimator to the pooled data.

In comparison with the main findings, for the alternative findings Central Bohemia is mostly estimated to be one of the most efficient regions. The results for Prague depend on the estimated variability of returns to scale. The non-parametric estimators allow high flexibility and thus Prague is estimated to be efficient. For the parametric estimators, Prague is inefficient because it lies relatively far behind the rest of the data cloud. For most of the methods, the Hradec Králové, Vysočina, Olomouc, and Zlín regions are among the less efficient regions.

A slight difference to the main finding is that most of the alternative methods find the Karlovy Vary region to be one of the most efficient, sometimes even super-efficient (meaning it needs to lie above the frontier). For the main methods, the efficiency of this region varies slightly but the region is positioned among the more efficient regions. Thus the difference does not seem to contradict the main findings and may potentially be attributed to methodological differences.

6.3 Effects on efficiency scores

Apart from the assessment of the relative efficiency of the Czech regions, this thesis aims to study the effects of selected conditions on the efficiency scores. As described in the section 5.4, the variables selected to represent these conditions include average age of residents, density of population, GDP per capita, index of political fractionalization, percentage of low-educated population, and separate percentage shares of communist and social democratic parties in the regional council. The effects on the efficiency scores are first studied for the main results. Then follows the summary of the effects for the alternative results.

Where feasible, the effects are analysed using the one-step analysis for the MLE estimation of SFA method. For the deterministic methods as well as for the results estimated by fixed or random effects, the one-step estimation is not possible. Thus even though they are not optimal, the results from the two-step estimation, using Tobit regression to determine effects on efficiency scores, are presented for comparison.[22] The results are mainly presented with Prague excluded from the Tobit regression (in case of two-step estimation) or accounted for specifically (in case of one step estimation). Prague has significantly different values of most of the analysed explanatory variables and hence it might shift the estimated effects if included. The effects with Prague are partially mentioned in the text for comparison.

For the main results the effects are studied primarily for the time-varying score (which is the same for KH95 and KLH14) and for the persistent score for KLH14 which both allow the one-step MLE estimation. The outcomes for the two-step Tobit regression for the time-invariant score (time-invariant efficiency for fixed effects and persistent efficiency for KH95) and for both total efficiency scores (for KH95 as well as for KLH14) are added for comparison. The signs and significance for the effects, using both the main and the alternative results, are presented in the table 6.2.

[22] For DEA the theoretical part of this thesis introduced a way to add exogenous inputs. Nevertheless not all of these conditions can be classified as exogenous inputs. Moreover, as we reasoned against large number of output variables, this would make the observations less comparable and thus the estimator would not be able to reject efficiency in most of the cases. Therefore two-step estimation is used for DEA as well.

For all the main results the density seems to have a negative effect, irrespective of the inclusion of Prague or exclusion of the Moravian-Silesian region (to account for the density outliers). This analysis originally argued that with lower density it might be more difficult and more costly to perform the services needed and thus density might have a positive effect. Nevertheless, the results are not in favour of this hypothesis and point in the opposite direction. Neither of the one-step estimators estimates the effect as significant. The two-step estimation provides a significantly negative effect for the time-invariant score and marginally significant for the total KH95 score. Therefore from the main results the effect of density cannot be clearly decided. The results point to a negative effect but the coefficients are mostly insignificant.

Similar to the density, coefficients on average age have consistently negative signs. This confirms the direction of the original hypothesis which claims that it might be more difficult to provide services to regions with higher average age. Additionally, this thesis argues that older population tends to prefer traditions over optimizing changes. While the effect is insignificant for the time-varying score, it significantly negatively affects the persistent efficiency from the KLH14 model when Prague is included. This is confirmed by significantly negative effect for the time-invariant efficiency score and for both total efficiency scores (both with and without Prague in the regression).

Most of the results claim that GDP per capita has a significantly positive effect on efficiency. In the discussion, arguments for both directions of the effects are found. From these results it seems that the regions do not care less about efficiency when they are wealthier. The reason for the positive effect seems to be that in regions with higher GDP per capita, there are on average richer people who may have a better financial education. Thus a richer region may imply more supervision of the public spending from the side of the population. To determine the causality and reject the suggestion that efficiency may influence the wealth of regions, a more thorough analysis would be needed. To indicate the result we tested for exogeneity of GDP in the two-step estimations of efficiency effects. The analysis using primary school students as instrument satisfies all the necessary tests and does not reject the hypothesis of exogeneity. This suggests no reverse causality between the efficiency scores and GDP per capita.

The exception to the significance of GDP is the one-step estimation of the time-varying score where the effect is estimated to be insignificant and even negative. However, the one-step estimation measures the effect on the variation of inefficiency[23] and thus the negative sign does not have to imply that GDP would cause inefficiency. In fact, the lowest (and most dispersed) time-varying efficiency scores correspond to the highest time-invariant efficiency scores (Central Bohemia and Prague). Therefore this does not contradict the findings of the other results which indicate that GDP has a positive effect on efficiency.

The fractionalization index is significant for almost all the main results. The sign is consistently positive for all the estimations which implies that higher number of political parties in the regional council have a positive effect on efficiency. As for GDP per capita, the effect of the political fractionalization is not clear from the theory. From the results it follows that more complicated discussion due to higher number of parties does not lead to inefficiency. Instead, the results support the hypothesis that with more political parties there is higher efficiency because there is more control of expenditure, while one party with a political majority can decide on its own about the expenditure. This agrees with the results of Palguta (2015).

An unexpected finding are the effects of the percentage of communists and social democrats in the regional council. For all the main results, both effect are estimated to be positive. In the one-step estimations the effect of the communist party is not significant. Due to discontinuities of the data, the effect of the social democratic party could be studied only for one of the one-step regressions but it is estimated significantly positive there. For all the main two-step estimations both of the effects are positive and significant.

There are a few hypotheses but no clear explanation for this finding. While the communist and social democratic parties may in fact lead regions more efficiently, it could be a result of a hidden correlation. The potential correlation with omitted variable cannot be rejected as it would require a complete analysis of the influences on efficiency. However similar finding is offered by De Borger & Kerstens (1996) who for Belgian municipalities registered a positive effect of the presence of socialist parties in governing coalitions.

The last to be discussed is the effect of the percentage of the population with lower than secondary education. In most of the investigations whether Prague is included or not, the effect is not found significant and does not have a consistent sign. The one-step analysis of the time-varying efficiency

[23]The one-step estimation investigates the effect of the explanatory variables on the variance of inefficiency, not on the individual scores. However, as the half-normal distribution of inefficiency is used, it can be deduced that a positive effect on variance means a negative effect on the estimated efficiency scores.

score estimates the effect as significantly negative. This is according to the expectations as it was argued that with less educated population there is less pressure on the regional governments to behave efficiently. Yet in some of the analyses the effect is estimated to be positive - for the time-invariant score it is even significantly positive. Thus based on this analysis, it cannot be clearly decided what effect the low-educated population has on efficiency. The prevailing insignificance of the effect may indicate that the effect is already captured by the GDP per capita as we argued that GDP per capita might be correlated with the better financial education of the population. The financial literacy may be more important than the total education.

For the alternative results, the results mostly support the findings of the main analysis. For all the non-parametric methods, the density and GDP per capita are significant. Similarly to the main results, GDP per capita is estimated to have a positive sign. Density has a significantly negative sign for all the non-parametric methods irrespectively of the exclusion of the two outliers for density. The sign is the same as for the main results and the significance might imply that some effect of density may exist although it has only weak support from the results of main analyses.

The rest of the explanatory variables in investigations of non-parametric results are mostly insignificant but all have consistent signs, which are same as for the main results with the exception of the ambiguous low-educated population. The average age and fractionalization have significant coefficients only for the DEA analysis, the percentage of communists and social democrats have significantly positive effects for both DEA and FDH analysis. The low-educated population is estimated to have insignificant coefficient for all the non-parametric methods.

The remaining methods provide various results. But up to a few exceptions the signs and significance is same as for the main results. The methods using COLS both for production and cost frontier produce similar outcomes to the results from the main and non-parametric methods, where the density and average age have a significantly negative effect on efficiency and GDP per capita and all political explanatory variables a significantly positive effect. The panel data analysis of the production frontier confirms the positive effect of GDP per capita and negative effect of average age. It produces a significantly negative effect of the percentage of low-educated population. Yet it suggests a positive sign for the effect of density as was originally suggested. Thus the effect of density is questionable although it points to a negative influence in most of the analyses.

The last method, pooled data analysis for the cost frontier using MLE, provides efficiency scores that are very similar to the time-varying scores for the main result. Therefore it is not surprising that it suggests a negative effect of GDP per capita. Due to the shape and location of the frontier, the poorest Karlovy Vary region is estimated to be the most efficient and the richest regions lag behind the frontier significantly. Apart from this contradictory finding, this method suggests a significantly positive effect of the fractionalization index.

To summarize the effects, based on our findings higher GDP per capita tends to have a positive effect on efficiency. Similarly, positive effect is estimated for higher number of political parties in the regional council but with higher shares for the communist and social democratic parties. The effects of average age of residents and density of population appear to be negative (although some counter-evidence for the negative effect of density is detected). Additionally, this analysis does not find any significant effect of the percentage of low-educated population on efficiency.

Table 6.2: Effects of the explanatory variables on the efficiency scores

method	density	aver. age	GDP p.c.	fraction.	low educ.	comm.	soc. dem.
time invar.	- ***	- **	+ ***	+ **	+ *	+ **	+ **
var KH95	-0	-0	-0	+ *	- **	+0	+ **
tot KH95	- *	- ***	+ ***	+ **	+0	+ **	+ **
cons KLH14	-0	-0	+ **	+0	+0	-0	X
tot KLH14	-0	- ***	+ ***	+ ***	+0	+ ***	+ ***
DEA	- ***	- *	+ ***	+ **	+0	+ ***	+ ***
FDH	- **	-0	+ ***	+0	+0	+ **	+ **
PFA	- ***	-0	+ ***	+0	+0	+0	+0
prod pool	- **	- ***	+ ***	+ ***	-0	+ ***	+ ***
prod FE	+ **	- ***	+ ***	+ *	- ***	+ ***	+ ***
cost pool COLS	- ***	- **	+ **	+ **	+0	+ ***	+ ***
cost pool MLE	+ **	-0	- **	+ *	-0	+ **	-0

Source: Author's calculations. *** denote significance at 1%, ** denote significance at 5%, * denotes significance at 10% level. 0 denotes insignificance. + and - stands for the signs of the effects which are included even for the insignificant coefficients. X is written in place of a coefficient that could not be estimated.

The results presented are compared for various estimators and in most cases the methods agree on the sign and significance of the effects. Nevertheless, the applications based on these findings must proceed carefully. As was already mentioned, to confirm some of the effects an exhausting analysis of the effects on efficiency would be needed. Furthermore, data used in this analysis may in some cases be suboptimal despite all the efforts employed.

The potential application of the results lies in the range of support and incentives for the regions that are estimated to be inefficient. Policies may be formed to repress adverse environment that affects the efficiency. Based on the positive effect of GDP per capita, financial literacy could be encouraged. Higher political competition may be supported to profit from the positive effect of fractionalization. To capitalize the finding that higher average age causes lower efficiency a specific analysis would need to be undertaken to determine the origin of the effect. Only after this analysis may a policy to counter this effect be formed. Similarly, further investigations of the reasons for the estimated positive effects of both left-wing political parties may be considered.

Chapter 7

Conclusion

The aim of the current analysis is to study the methods of efficiency analysis at first theoretically and later to apply these to Czech regional data. Various competing and complementary methods are considered. All the techniques considered are defined and discussed. Their common extensions are explained. Panel data applications are discussed for all the methods. The discussion of the efficiency estimators further includes comparison of the methods in general and particularly for the efficiency analysis of public spending.

Despite the general complementarity of the methods, several methods are preferred for the public efficiency analysis. Some theoreticians advocate the use of the FDH method which has the least assumptions if the deterministic feature is not taken as an assumption. It corresponds to the ambiguity of theory about the "production process" of the public sector. On the other hand when inputs are not clear and costs are used in place of input, the application of the cost frontier approach of the SFA method is suggested to utilize the form of the data. This approach allows a simple use of more outputs, similarly as for the non-parametric methods. Yet both for the non-parametric and parametric methods, the use of more outputs and more inputs is limited corresponding to the number of units used for the analysis. Further advantage of the cost frontier approach is that the SFA techniques are more flexible for panel data analyses. Thus for the efficiency analysis of public spending the cost frontier approach of SFA method seems to be the most convenient.

The main analysis of this thesis applies the cost frontier approach of SFA to the data on Czech regions in years 2003–2014. The resulting efficiency scores are provided and the relative inefficiencies are discussed. The most efficient region seems to be Central Bohemia. It holds its prime position for most of the years when the efficiency is allowed to vary over time. To the next most efficient regions belong Karlovy Vary and South Moravia regions and, depending on the choice of method, the capital city Prague. Prague is considered an outlier with unmatchingly high expenditure and slightly higher score of the output index. Thus its efficiency depends on the flexibility and assumptions on returns to scale incorporated in the estimation methods. The least efficient regions are in almost every case Olomouc and Zlín regions. These findings hold for most of the methods used in the robustness study, where the main results are compared to the outcomes of non-parametric and various parametric estimators.

The efficiency scores vary over time when the time-invariance assumption is dropped. Most significant improvements occur for the efficiency scores of Prague and Central Bohemia which in the main analysis reach their prime position in later years. Two of the main models differ in the separation of permanent efficiency scores from random heterogeneity effect. As their results differ partially, the efficiency appears to be distinguishable from the individual heterogeneity. The model that separates heterogeneity and constant efficiency provides less disperse inefficiency estimates. The average score is relatively high (0.882). This model also favours some of the regions that have adverse environment but might otherwise perform efficiently, as for example the Ústí nad Labem region.

The aim of this analysis is not only to rank the regions and discuss the relative efficiency and its evolution over time. To benefit from the analysis practically, several exogenous conditions are investigated for their potential effect on efficiency. Wealth, measured by GDP per capita, appears to have a positive effect on efficiency. Both density of population and average age of regional residents tend to influence the efficiency negatively. The index of political fractionalization, indicating whether a region is ruled by a single party or by multiple parties, points to a positive effect of multiple-party council on efficiency. It shows that with higher number of political parties, population may benefit from the political compromises and better control. Similarly, a significantly positive effect is found for higher shares of both communist and social democratic parties. No effect is proved for higher share of low-educated residents.

This thesis does not aim to undertake an exhausting analysis of effects on efficiency. Only selected conditions are considered. The complete investigation is left to future researchers along with the confirmation of the currently estimated effects by considering wider spectre of causes. Further analyses could also consider only particular field of regional services and investigate whether the effects hold generally or only for a specific section of regional governance.

This analysis may suffer from some limitations. To the limitations inherent to public efficiency analyses belong low availability of appropriate data and unclear theoretical specification of inputs and outputs for public analyses. The data used in the analysis are selected as the best available for the range of observations. Input and output indicators are selected correspondingly to previous analyses after a detailed investigation of regional services. Nevertheless, the analysis may be repeated when more data become available.

Bibliography

Afonso, A. & S. Fernandes (2006): "Measuring local government spending efficiency: evidence for the Lisbon region." *Regional Studies* **40(1)**: pp. 39–53.

Aigner, D., C. K. Lovell, & P. Schmidt (1977): "Formulation and estimation of stochastic frontier production function models." *Journal of Economics* **6(1)**: pp. 21–37.

Aigner, D. J. & S. F. Chu (1968): "On estimating the industry production function." *The American Economic Review* pp. 826–839.

Anderson, T. W. (1951): "Estimating linear restrictions on regression coefficients for multivariate normal distributions." *Annals of Mathematical Statistics* **22**: pp. 327–51.

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

Banker, R. D. & R. C. Morey (1986): "Efficiency analysis for exogenously fixed inputs and outputs." *Operations Research* **34(4)**: pp. 513–521.

Charnes, A., W. W. Cooper, & E. Rhodes (1978): "Measuring the efficiency of decision making units." *European Jounal of Operational Research* **2(6)**: pp. 429–444.

Charnes, A., W. W. Cooper, & E. Rhodes (1981): "Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through." *Management science* **27(6)**: pp. 668–697.

Coelli, T. J., D. S. P. Rao, C. J. O'Donnell, & G. E. Battese (2005): *An introduction to efficiency and productivity analysis.* Springer Science & Business Media.

Cooper, W. W., L. M. Seiford, & K. Tone (editors) (2007): *Data Envelopment Analysis: A Comprehensive text with models, applications, references, and DEA-solver software.* Boston: Kluwer Academic Publishers.

Cooper, W. W., L. M. Seiford, & J. Zhu (2011): "Data Envelopment Analysis: History, models, and interpretations." In W. W. Cooper, L. M. Seiford, & J. Zhu (editors), "Handbook on Data Envelopment Analysis," volume 164, pp. 1–39. Springer Science & Business Media.

Cornwell, C., P. Schmidt, & R. C. Sickles (1990): "Production frontiers with cross-sectional and time-series variation in efficiency levels." *Journal of econometrics* **46(1)**: pp. 185–200.

Daraio, C. & L. Simar (2007): *Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications*, volume 4. Springer Science & Business Media.

De Borger, B. & K. Kerstens (1996): "Cost efficiency of Belgian local governments: A comparative analysis of FDH, DEA, and econometric approaches." *Regional Science and Urban Economics* **26(2)**: pp. 145–170.

De Borger, B., K. Kerstens, W. Moesen, J. Vanneste *et al.* (1994): "A non-parametric free disposal hull (FDH) approach to technical efficiency: an illustration of radial and graph efficiency measures and some sensitivity results." *Swiss Journal of Economics and Statistics* **130(4)**: pp. 647–667.

Debreu, G. (1951): "The coefficient of resource utilization." *Econometrica: Journal of the Econometric Society* pp. 273–292.

Deprins, D., L. Simar, & H. Tulkens (1984): "Measuring labor inefficiency in post offices." *The Performance of Public Enterprises: Concepts and measurements* pp. 243–267.

Färe, R., S. Grosskopf, J. Logan, & C. K. Lovell (1985): "Measuring efficiency in production: with an application to electric utilities." In "Managerial Issues in Productivity Analysis," pp. 185–214. Springer.

Farrell, M. J. (1957): "The measurement of productive efficiency." *Journal of the Royal Statistical Society. Series A (General)* **120(3)**: pp. 253–290.

Johnson, A. L. & T. Kuosmanen (2015): "An introduction to CNLS and StoNED methods for efficiency analysis: Economic insights and computational aspects." In S. C. Ray, S. C. Kumbhakar, & P. Dua (editors), "Benchmarking for Performance Evaluation," pp. 117–186. Springer.

Jondrow, J., C. K. Lovell, I. S. Materov, & P. Schmidt (1982): "On the estimation of technical inefficiency in the stochastic frontier production function model." *Journal of econometrics* **19(2)**: pp. 233–238.

Kalb, A. (2010): *Public Sector Efficiency: Applications to Local Governments in Germany*. Springer Science & Business Media.

Kneip, A., B. U. Park, & L. Simar (1998): "A note on the convergence of nonparametric DEA estimators for production efficiency scores." *Econometric theory* **14(06)**: pp. 783–793.

Koopmans, T. C. (editor) (1951): *Activity Analysis of Production and Allocation*. New York: Wiley.

Kumbhakar, S. C. & A. Heshmati (1995): "Efficiency measurement in Swedish dairy farms: an application of rotating panel data, 1976–88." *American Journal of Agricultural Economics* **77(3)**: pp. 660–674.

Kumbhakar, S. C., G. Lien, & J. B. Hardaker (2014): "Technical efficiency in competing panel data models: a study of Norwegian grain farming." *Journal of Productivity Analysis* **41(2)**: pp. 321–337.

Kumbhakar, S. C., H. Wang, & A. P. Horncastle (2015): *A Practitioner's Guide to Stochastic Frontier Analysis Using Stata*. Cambridge University Press.

Kumbhakar, S. C. & H.-J. Wang (2005): "Estimation of growth convergence using a stochastic production frontier approach." *Economics Letters* **88(3)**: pp. 300–305.

Kuosmanen, T., A. Johnson, & A. Saastamoinen (2015): "Stochastic nonparametric approach to efficiency analysis: a unified framework." In J. Zhu (editor), "Data Envelopment Analysis," pp. 191–244. Springer.

Mandl, U., A. Dierx, & F. Ilkovitz (2008): "European economy. the effectiveness and efficiency of public spending. economic and financial affairs, european commission." *Economic Papers* **301**.

Murillo-Zamorano, L. R. (2004): "Economic efficiency and frontier techniques." *Journal of Economic surveys* **18(1)**: pp. 33–77.

Palguta, J. (2015): "Political rent-seeking in public procurement: Evidence from the entry of political challengers at electoral thresholds." *CERGE-EI Working Paper Series* **(549)**.

Park, B. U., L. Simar, & C. Weiner (2000): "The FDH estimator for productivity efficiency scores." *Econometric Theory* **16(06)**: pp. 855–877.

Rae, D. (1968): "A note on the fractionalization of some european party systems." *Comparative Political Studies* **1(3)**: p. 413.

Ray, S. C. & L. Chen (2015): "Data envelopment analysis for performance evaluation: A child's guide." In S. C. Ray, S. C. Kumbhakar, & P. Dua (editors), "Benchmarking for Performance Evaluation," pp. 75–116. Springer.

Simar, L. & P. W. Wilson (1998): "Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models." *Management science* **44(1)**: pp. 49–61.

Simar, L. & P. W. Wilson (2011): "Performance of the bootstrap for DEA estimators and iterating the principle." In W. W. Cooper, L. M. Seiford, & J. Zhu (editors), "Handbook on Data Envelopment Analysis," volume 164, pp. 241–271. Springer Science & Business Media.

Stock, J. H. & M. Yogo (2005): "Testing for weak instruments in linear IV regression." *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* .

Vanden Eeckhaut, P., H. Tulkens, & M. Jamar (1993): "Cost efficiency in Belgian municipalities." In C. Fried, S. Lovell, & Schmidt (editors), "The Measurement of Productive Efficiency: Techniques and Applications.", Oxford University Press, New York.

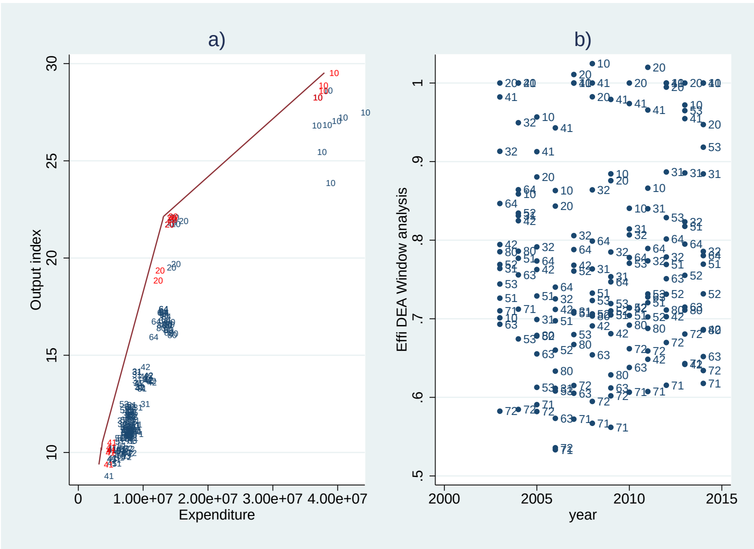
Worthington, A. C. (2000): "Cost efficiency in australian local government: A comparative analysis of mathematical programming anf econometrical approaches." *Financial Accountability & Management* **16(3)**: pp. 201–223.

Appendix A

Results from alternative models

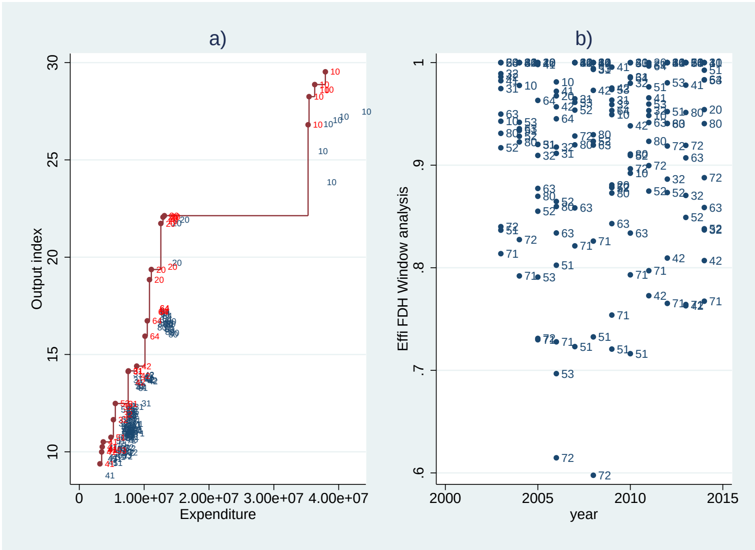
This appendix presents the results from the alternative models that are compared to the main results in section 6.2.

Figure A.1: Results using DEA estimator



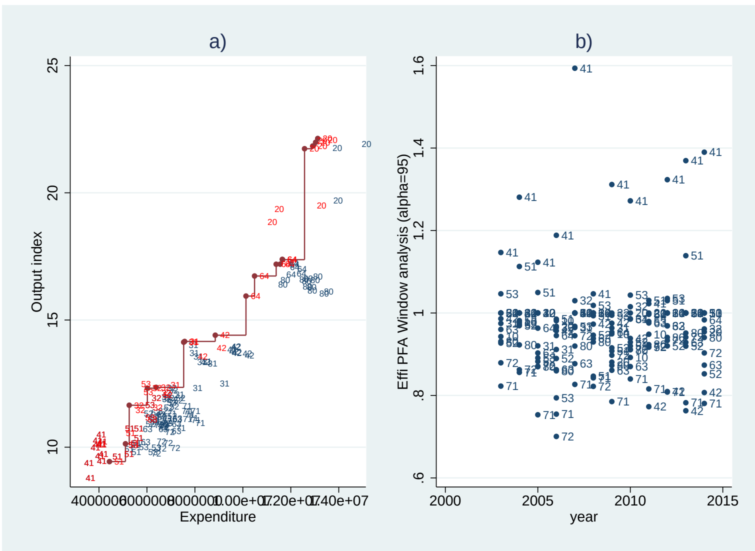
Source: Author's computations. *Note:* The regions are denoted in the following way: 10 = capital city Prague, 20 = Central Bohemian, 31 = South Bohemian, 32 = Plzeň, 41 = Karlovy Vary, 42 = Ústí nad Labem, 51 = Liberec, 52 = Hradec Králové, 53 = Pardubice, 63 = Vysočina, 64 = South Moravian, 71 = Olomouc, 72 = Zlín, 80 = Moravian-Silesian.

Figure A.2: Results using FDH estimator



Source: Author's computations. *Note:* The regions are denoted in the following way: 10 = capital city Prague, 20 = Central Bohemian, 31 = South Bohemian, 32 = Plzeň, 41 = Karlovy Vary, 42 = Ústí nad Labem, 51 = Liberec, 52 = Hradec Králové, 53 = Pardubice, 63 = Vysočina, 64 = South Moravian, 71 = Olomouc, 72 = Zlín, 80 = Moravian-Silesian.

Figure A.3: Results using PFA estimator

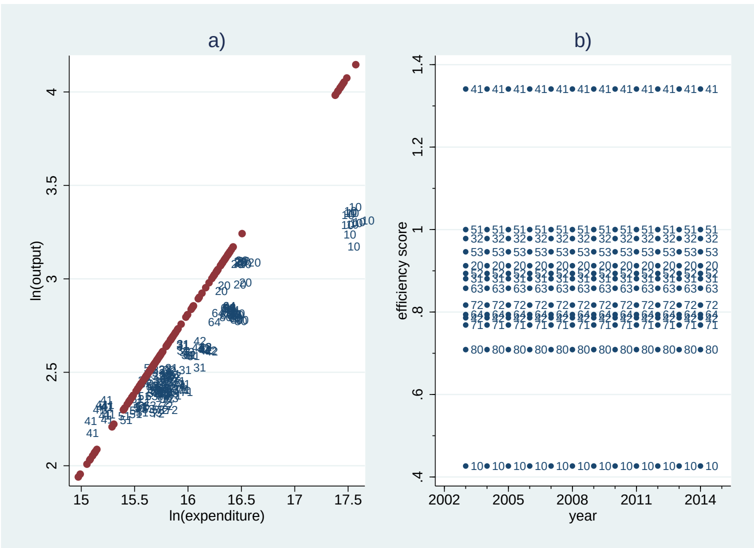


Source: Author's computations. *Note:* The regions are denoted in the following way: 10 = capital city Prague, 20 = Central Bohemian, 31 = South Bohemian, 32 = Plzeň, 41 = Karlovy Vary, 42 = Ústí nad Labem, 51 = Liberec, 52 = Hradec Králové, 53 = Pardubice, 63 = Vysočina, 64 = South Moravian, 71 = Olomouc, 72 = Zlín, 80 = Moravian-Silesian. Prague is not included in the graph of frontier, figure (a), to allow better visibility of the super-efficient observations, lying above frontier.

Figure A.4: Results using SFA estimator - production frontier (pooled data - COLS)

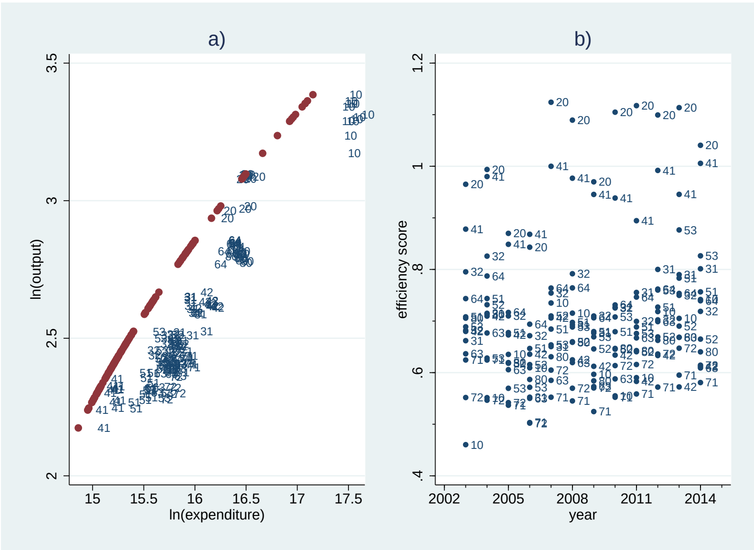
Source: Author's computations. *Note:* The regions are denoted in the following way: 10 = capital city Prague, 20 = Central Bohemian, 31 = South Bohemian, 32 = Plzeň, 41 = Karlovy Vary, 42 = Ústí nad Labem, 51 = Liberec, 52 = Hradec Králové, 53 = Pardubice, 63 = Vysočina, 64 = South Moravian, 71 = Olomouc, 72 = Zlín, 80 = Moravian-Silesian.

Figure A.5: Results using SFA estimator - production frontier (panel data - fixed effects)



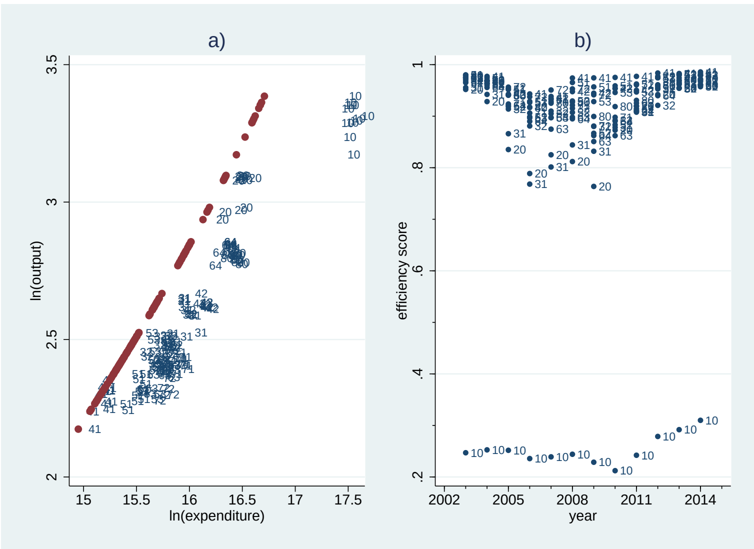
Source: Author's computations. *Note:* The regions are denoted in the following way: 10 = capital city Prague, 20 = Central Bohemian, 31 = South Bohemian, 32 = Plzeň, 41 = Karlovy Vary, 42 = Ústí nad Labem, 51 = Liberec, 52 = Hradec Králové, 53 = Pardubice, 63 = Vysočina, 64 = South Moravian, 71 = Olomouc, 72 = Zlín, 80 = Moravian-Silesiann.

Figure A.6: Results using SFA estimator - cost frontier (pooled data - COLS)



Source: Author's computations. *Note:* The regions are denoted in the following way: 10 = capital city Prague, 20 = Central Bohemian, 31 = South Bohemian, 32 = Plzeň, 41 = Karlovy Vary, 42 = Ústí nad Labem, 51 = Liberec, 52 = Hradec Králové, 53 = Pardubice, 63 = Vysočina, 64 = South Moravian, 71 = Olomouc, 72 = Zlín, 80 = Moravian-Silesian.

Figure A.7: Results using SFA estimator - cost frontier (pooled data - MLE)



Source: Author's computations. *Note:* The regions are denoted in the following way: 10 = capital city Prague, 20 = Central Bohemian, 31 = South Bohemian, 32 = Plzeň, 41 = Karlovy Vary, 42 = Ústí nad Labem, 51 = Liberec, 52 = Hradec Králové, 53 = Pardubice, 63 = Vysočina, 64 = South Moravian, 71 = Olomouc, 72 = Zlín, 80 = Moravian-Silesian.