Department of Information and Service Economy

Modelling heterogeneous operating environment and production risk in modern productivity analysis

Antti Saastamoinen

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 119/2014

Modelling heterogeneous operating environment and production risk in modern productivity analysis

Antti Saastamoinen

Aalto University
School of Business
Department of Information and Service Economy

Supervising professor
Professor Timo Kuosmanen

Opponent
Emeritus Professor Thomas Weyman-Jones, Loughborough University, United Kingdom

Aalto University publication series
DOCTORAL DISSERTATIONS 119/2014

ISBN 978-952-60-5808-5
ISBN 978-952-60-5809-2 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5809-2

Images: Pohjolan Sanomat, photographer Nina Susi (cover)

Unigrafia Oy
Helsinki 2014

Finland

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Antti Saastamoinen

Name of the doctoral dissertation
Modelling heterogeneous operating environment and production risk in modern productivity analysis

Publisher School of Business

Unit Department of Information and Service Economy

Series Aalto University publication series DOCTORAL DISSERTATIONS 119/2014

Field of research Productivity and efficiency analysis

Date of the defence 26 September 2014

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

This dissertation studies the implications of a risky operating environment of firms for productivity analysis. The discussion is done under a new methodological framework and the conceptual gap between the models of production risk and the models of inefficiency is bridged. The thesis includes an introductory part and five research articles.

Article 1 introduces the new stochastic semi-nonparametric envelopment of data (StoNED) estimation method that has been used in the thesis.

Article 2 applies the StoNED method in the context of Finnish electricity distribution regulation. In 2010, the Finnish regulatory agency adopted StoNED to assess the cost efficiency of distribution companies. We compare StoNED to the previously used methods and it outperforms them because it is more flexible in taking into account the firm characteristics and its environment.

Article 3 is a literature review which explores the links between production risk and inefficiency literature. It is found that the econometric concept of heteroscedasticity connects these two veins of literature. However, combining risk and the models of inefficiency may be complicated as it is challenging to distinguish the effects of inefficiency from the effects of the risky environment.

In article 4, we examine the relationship between macro-productivity and corruption. Conventionally corruption is seen either as an impediment to or, under certain conditions, as a catalyst of economic growth. We, however, consider that corruption acts as a macro risk factor. That is, corruption increases the likelihood of productivity being low, but it allows the possibility of good performance also. Our analysis shows that the dispersion of productivity is larger among countries with high corruption levels.

Article 5 studies the quality of service of Finnish electricity distribution firms in terms of costs of interruptions. We study how the underground cabling affects these costs and how the quality targets should be set. As expected, underground cabling decreases the level of costs. However, underground cabling does not significantly decrease the variation in these costs. We observe rather large variation of interruption costs even at relatively high degrees of underground cabling because the interruptions are costly in the underground networks mainly located in big cities. We also suggest that the quality targets should be set by using a StoNED based on a best-practice quality frontier rather than by an average of past performances of firms. The frontier is better for incentivizing firms to improve their operations and it produces more balanced targets than the average.

Keywords operating environment, risk, productivity, efficiency

ISBN (printed) 978-952-60-5808-5	**ISBN (pdf)** 978-952-60-5809-2	
ISSN-L 1799-4934	**ISSN (printed)** 1799-4934	**ISSN (pdf)** 1799-4942
Location of publisher Helsinki	**Location of printing** Helsinki	**Year** 2014
Pages 237	**urn** http://urn.fi/URN:ISBN:978-952-60-5809-2	

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Antti Saastamoinen

Väitöskirjan nimi
Heterogeenisen tuotantoympäristön ja tuotantoriskin mallintaminen modernissa tuottavuusanalyysissa

Julkaisija Kauppakorkeakoulu

Yksikkö Tieto- ja palvelutalouden laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 119/2014

Tutkimusala Tuottavuus- ja tehokkuusanalyysi

Väitöspäivä 26.09.2014

☐ **Monografia** ☒ **Artikkeliväitöskirja** ☐ **Esseeväitöskirja**

Tiivistelmä

Tämä väitöskirja tutkii yritysten toimintaympäristön riskien vaikutuksia tuottavuusanalyysille. Aihepiiriä lähestytään uuden empiirisen menetelmän kautta. Käsitteellisellä tasolla aiemmassa tutkimuksessa varsin erillisinä esiintyneet tuotantoriskien ja tehottomuuden mallit tuodaan myös lähemmäksi toisiaan. Tämä väitöskirja koostuu esittelyosasta ja viidestä tutkimusartikkelista.

Artikkeli 1 esittelee StoNED estimointimenetelmän, jota väitöskirjassa sovelletaan.

Artikkelissa 2 StoNED menetelmää sovelletaan suomalaisten sähkönjakeluverkkoyhtiöiden hintaregulaation yhteydessä. Vuonna 2010 Suomen energiamarkkinoiden valvontaviranomainen otti StoNED menetelmän käyttöön arvioidakseen yhtiöiden kustannustehokkuutta. Vertaamme StoNED menetelmää aiemmin käytettyihin menetelmiin. Havaitsemme, että StoNED on soveltuvampi menetelmä sillä se ottaa paremmin huomioon yhtiöiden ominaisuudet ja toimintaympäristön.

Artikkeli 3 on kirjallisuuskatsaus, joka tutkii tuotantoriskeihin ja tehottomuuteen keskittyvien kirjallisuudenalojen yhteyksiä. Artikkelissa osoitetaan, että ekonometriasta tuttu heteroskedastisuuden käsite yhdistää nämä kirjallisuuden alat. Menetelmällisesti käsitteiden yhdistäminen on kuitenkin haastavaa, sillä on vaikeaa erottaa toimintaympäristön riskien vaikutuksia aidosta tehottomuudesta.

Artikkelissa 4 tutkimme makrotason tuottavuuden ja korruption yhteyttä. Tyypillisesti kirjallisuudessa korruptio on nähty joko vähentävän, tai tietyissä olosuhteissa parantavan taloudellista kasvua. Tämä tutkimus kuitenkin näkee korruption makroriskinä, tarkoittaen, että korruptio lisää todennäköisyyttä heikkoon tuottavuuteen, mutta mahdollistaen myös hyvän tuottavuuden tason. Analyysi osoittaa, että tuottavuuden hajonta on suurempaa korruption kasvaessa.

Artikkelissa 5 tutkimme suomalaisten sähkönjakeluverkkoyhtiöiden palvelun laatua keskeytyskustannuksien kautta. Tutkimme kuinka maakaapelointi vaikuttaa keskeytyskustannuksiin ja kuinka yhtiöiden laatutavoitteet tulisi asettaa. Odotetusti suurempi maakaapelointiaste vähentää keskeytyskustannuksia. Maakaapelointi ei kuitenkaan merkitsevästi vähennä näiden kustannusten hajontaa. Kustannusten hajonta on huomattavaa varsin suurillakin maakaapelointiasteilla, koska keskeytykset suurien kaupunkien maakaapeliverkoissa ovat kalliita. Ehdotamme myös, että laatutavoitteet tulisi asettaa StoNED menetelmällä estimoituun parhaisiin toimintatapoihin perustuvan laaturintaman avulla yhtiöiden oman menneen laatutason keskiarvon sijaan. Laaturintama kannustaa paremmin yhtiöitä parantamaan toimintaansa ja se tuottaa tasapuolisempia tavoitteita kuin keskiarvo.

Avainsanat toimintaympäristö, riski, tuottavuus, tehokkuus

ISBN (painettu) 978-952-60-5808-5	**ISBN (pdf)** 978-952-60-5809-2	
ISSN-L 1799-4934	**ISSN (painettu)** 1799-4934	**ISSN (pdf)** 1799-4942
Julkaisupaikka Helsinki	**Painopaikka** Helsinki	**Vuosi** 2014
Sivumäärä 237	**urn** http://urn.fi/URN:ISBN:978-952-60-5809-2	

Acknowledgements

On a path towards a PhD, everyone must walk with their own feet. But surely nobody has walked that path alone. Indeed, although only my name reads on the cover of this dissertation, it would be hugely wrong to interpret it as though no one else has been involved in my process of obtaining a PhD. Over the years, I have been fortunate enough to gain friendship, support, encouragement, advice and guidance from many people, both in academia and outside of it. I am deeply grateful for all of your time, kindness, and patience during these years.

First, I most gratefully thank Professor Timo Kuosmanen, who saw the researcher potential in me at the time when I personally was somewhat losing my own faith in that potential. His active take on supervising my PhD has encouraged me to keep going, and it has been a pleasure to work with him. Timo gave me freedom to act as an individual researcher but also was there set out clear objectives for my PhD project at times when I had doubts about its direction. Besides doing good research, I have also learned much about the workings of academia from him. His advices I will remember throughout my career.

I would also like to thank two other co-authors of mine, Associate Professor Andrew L. Johnson from Texas A&M University and Dr. Timo Sipiläinen from University of Helsinki. The first time I met Professor Johnson was in my first conference in 2010, and since then I have been privileged to be able to take part in the lively discussions with him and Professor Kuosmanen. The opportunity to listen to the discussions between these two professors has taught me things about productivity and efficiency analysis that I probably would not have learned from any books or journal articles. I have also enjoyed the discussions with Dr. Sipiläinen who has always been kindly asking about my progress in my studies every time we have met.

Both Professor Jyrki Wallenius and Emeritus Professor Pekka Korhonen have been instrumental in giving me the opportunity to work at the Department of Information and Service Economy and also in providing

the necessary funding for me to pursue my PhD. Their positive attitude towards the whole work community has created an inspiring working environment at the department. I thank them for the opportunity to be a part of the department.

A number of colleagues and friends at the Department of Information and Service Economy deserve special acknowledgements. First, I would like to thank Oskar Ahlgren, with whom I have had the pleasure to share an office for the major part of my stay at the department. I am glad to say that the initial collegial relationship has turned into a good friendship during these years. Dr. Abolfazl Keshvari and Nasim Dehghan I consider the kindest people I have ever met, and I am glad that I have had the chance to get to know them, not only as colleagues but as friends also. I give my thanks also to Anton Frantsev, Yulia Tammisto, and Bikesh Upreti. The time spent with you either at the office or outside of it has always lifted my mood so that it has been easier to think about the thesis. You are a proof that the department has surely given me many new good friends. I also would like to thank Dr. Ankur Sinha and Assistant Professor Pekka Malo. You both have been a good example for me, and I admire the level of professionalism and dedication that you show around the office. I shared an office with Ankur in the beginning of my PhD studies, and it truly inspired me for my PhD to see such a talent at work close by. Juha Eskelinen also set an example for me. Juha completed his PhD in spring 2014, and it was truly helpful to follow his last steps towards the PhD. Over the years, Juha has also provided many interesting insights in our discussions, since we worked in the same field of research. Lastly, I want to express my thanks to the whole administrative staff of the department, especially to Merja Mäkinen and Jutta Heino. Without your help, I would not have survived through all the practicalities related to my studies and research.

There are many people outside the immediate workplace to whom I want to express my gratitude for being a part of my trip towards the PhD. Sami Pakarinen, a fellow doctoral candidate, who started his PhD studies at the same time as I did many years back. Since then we have been friends even though our roads towards a PhD have taken somewhat different turns over the years. We have had many good conversations about work, research,

and life in general. I also have the pleasure to be the godfather of Sami's daughter. I thank him for this honour.

During the PhD process I was also involved in the board activities of the post graduate student association of Aalto University, Aallonhuiput. Many thanks go to all fellow board members with whom I have had many lively discussions about the trials and tribulations of doing research and life in academia.

Mika Kortelainen from the Government Institute of Economic Research has also had a special role in my life as a researcher. Many years back, when I was still studying in Joensuu, Mika was in fact giving me my first course in econometrics. I thank him for the job well done since his good take on teaching that course sparked my enthusiasm for studying economic phenomena with statistical methods. Later, our paths crossed again when I moved to Helsinki and I am glad that it happened as Mika has always been a good example for me as a researcher.

I also want to thank Professor Mikael Linden from University of Eastern Finland. He is responsible for getting me on the road of pursuing a PhD. I indeed started my PhD studies under his supervision back in 2008. Although life led me to go for a PhD here in Helsinki, I owe great thanks to Professor Linden since he saw such value in my Master's thesis that I was chosen for PhD studies.

Naturally, I also owe my deep gratitude to the two pre-examiners of my thesis, Professor Tom Weyman-Jones from Longborough University and Professor Rauli Svento from the University of Oulu. I appreciate the effort and time that they have put forward to examine my thesis. I thank them for all the feedback that they gave on my dissertation.

Up to this point, I have acknowledged people whom I have had the pleasure to meet during my PhD career. However, there are a number of people outside the academic circles whom I want to give my special thanks to.

First, I would like to thank Jarmo Maaranen and Janne Sormunen, my two long-time friends. I have known Jarmo for over fifteen years, and Janne even longer, straight from elementary school. I am glad that even though during the last six years or so, life has scattered us all around Finland, our friendship has not faded. I would also like to thank all other friends in

Joensuu. Each time I visit Joensuu, your company has offered me an escape from the somewhat hectic PhD life in Helsinki.

From family and relatives I first would like to thank my grandmother Enni and my late grandfather Erkki. They always supported us, grandchildren, and given their advice on life to us. I think that although we, grandchildren, may not realize that all the time, we have gained so much from them.

I also thank my aunt Sirpa. Although we see too little of each other these days, every time we meet, she has always been a pleasant companion to talk with about everything. I thank her for these discussions. I also want to give my thanks to two older cousins of mine, Jukka and Titta. Throughout my life, both have been good friends and examples for me.

Lastly and most importantly, I thank my mother Hilkka. There are probably not enough words to express my gratitude to her for all the work she has done to raise me up to be the man I am today. You have been there for me always when I have needed you, and I am eternally grateful to you. The times have been difficult every now and then, but we have managed to overcome them. I would not have gone through this PhD process without the perseverance that I learned from you.

Helsinki, June 16, 2014
Antti Saastamoinen

Contents

List of abbreviations

CE	Cost Efficiency
CES	Constant Elasticity of Substitution
CNLS	Convex Nonparametric Least Squares
COLS	Corrected OLS
CRS	Constant Returns to Scale
DEA	Data Envelopment Analysis
FDH	Free Disposal Hull
FGLS	Feasible Generalized Least Squares
GLS	Generalized Least Squares
ML	Maximum Likelihood
MOLS	Modified OLS
MoM	Method of Moments
OLS	Ordinary Least Squares
PPS	Production Possibility Set
SFA	Stochastic Frontier Analysis
StoNED	Stochastic semi-Nonparametric Envelopment of Data
TE	Technical Efficiency
VRS	Variable Returns to Scale

List of research articles

This doctoral dissertation consists of a summary and the following research articles which are referred to in the text by their numerals.

Article 1
Timo Kuosmanen, Andrew Johnson, and Antti Saastamoinen
Stochastic nonparametric approach to efficiency analysis: A Unified Framework, unpublished manuscript (Forthcoming in J. Zhu (Ed.) *Handbook on DEA*, Springer).

Article 2
Timo Kuosmanen, Antti Saastamoinen, and Timo Sipiläinen
What is the best practice for benchmark regulation of electricity distribution? Comparison of DEA, SFA and StoNED methods
Energy Policy (2013), vol. 61, pp. 740-750.
DOI: 10.1016/j.enpol.2013.05.091

Article 3
Antti Saastamoinen
Heteroscedasticity or Production Risk? A Synthetic View
Forthcoming in *Journal of Economic Surveys*, published online
DOI: 10.1111/joes.12054

Article 4
This is the authors accepted manuscript of an article published as the version of record in Applied Economics May 2014

Antti Saastamoinen and Timo Kuosmanen
Is Corruption Grease, Grit, or a Gamble? Corruption Increases Variance of Productivity Across Countries. *Applied Economics* (2014), vol. 46, Issue 23, pp. 2833-2849.
http://www.tandfonline.com/doi/full/10.1080/00036846.2014.914149

Article 5
Antti Saastamoinen and Timo Kuosmanen
Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland
Forthcoming in *Energy Economics*, published online.
DOI: 10.1016/j.eneco.2014.04.016

PART I: Overview of the dissertation

1. Background of the thesis

The traditional economic theory highlights that all firms operate efficiently.[1] For example, the most common behavioural assumption imposed on firms, profit maximization, implies that with given prices, firms produce the highest possible output from given inputs and that they do this with the lowest possible cost (Mas-Colell, Whinston, and Green, 1995). In terms of production function, this means that the production function applies to a firm when it is producing the maximum possible output from its inputs, which implies that its productivity is maximized. Thus, if we interpret the traditional theory very narrowly, all firms observed in the markets are technically efficient since inefficient firms are ultimately driven out from the markets. Obviously this is not the case. We do observe inefficiency and large productivity differences among firms. For example Syverson (2011) points out that persistent productivity differences are present virtually in all types of industries.[2] Syverson examines the determinants of productivity differences in detail and roughly categorizes them to intra-firm determinants and external determinants. The latter are related to the market conditions or, more generally, to the operation environment of firms. The former, on the other hand, relate to managerial talent, learning and R&D, among others. Over these factors firms usually have some control. As the following discussion will make it evident, this thesis is more concerned about the external influences of operating environment on productivity.

In order to make some practical judgements about the level of productivity, the amount by which productivity can be improved should be quantified. Indeed, it would be very hard for firm managers to make decisions concerning productivity if they were unaware about the productivity target that they should obtain. Identifying this target effectively

[1] For the sake of simplicity, we use the term 'firm' to refer to any entity that engages in production activities. The discussion and the methods applied in this thesis extend themselves for example to countries and public service providers which do not directly fall under the typical interpretation of a firm.

[2] Syverson uses the terms *productivity* and *efficiency* interchangeably. However, to be exact, (technical) efficiency is just one component of productivity (change), which includes also technical change and scale components, which account for productivity change due to change in technology and deficit in production, due to suboptimal scale size. But of course, lower efficiency means lower productivity.

means that the firm's operations should be compared to some ideal technology which describes the optimal way of production.

However, such theoretical ideals hardly exist, except maybe in engineering, and the technology has to be estimated from the observed data. Thus what is actually done is a comparison to the *best observed practices* of the industry. The conventional approach is to estimate a production or a cost function via the usual linear regression methods. Unfortunately, these methods do not explicitly acknowledge the presence of technical inefficiency as they are constructed upon the traditional economic theory. Firms are still assumed to succeed in their optimization of production in terms of technical efficiency. Although conventional empirical models do allow deviations from optimal production, these models usually downgrade these deviations simply as a statistical error without much interest to the analyst. That is, the resulting residual is seen only as an estimation error, and the interest is in the parameters of production function itself (see the discussion in Kuosmanen & Fosgerau, 2009). In cases where the residual is considered interesting, it is then lumped as a single productivity measure without further considerations of its content (see e.g. Abramowitz, 1956; Syverson, 2011). The interest of further research has then mainly been to study the factors that explain variations in this residual and not so much the magnitude of the residual.

Nevertheless, it is problematic if we do not have a very clear idea about what the residual itself actually contains. Through conventional modelling, it is practically impossible to identify the size of the residual part attributed to technical inefficiency and for example to measurement errors or specification errors of the model. This is fine if the interest indeed lies only in the production function parameters. But for managerial decisions, it would be important to explicitly identify the measure of inefficiency out of that residual. This has led to the development of methods that allow us to explicitly model this inefficiency.

Since the end of 1970's, the field of productivity and efficiency analysis has seen a surge of applications of so-called frontier methods, namely nonparametric data envelopment analysis (DEA) and parametric stochastic frontier analysis (SFA). Grounded on the path-breaking work of Farrell (1957), these methods were developed by Charnes et al. (1978; DEA) and Aigner et al. (1977; SFA) to estimate the efficient (production) technology of

firms under the presence of technical inefficiency. This allowed the analyst to compare the performance of individual firms against technology that is efficient and to assess the possible magnitude of inefficiency. Over the years, the application areas of these methods have ranged from the micro level to the aggregate macro level. In their survey of the literature, Fried et al. (2008) identified around 50 different areas in which these methods have been applied. One especially prominent field of applications, a field relevant also in this thesis, has been the performance of public services such as utilities (see e.g. Coelli and Lawrence, 2006). For example, many countries use these methods to incentivize electricity distribution companies to operate cost-efficiently. Otherwise, these companies are little incentivized to act so, due to their natural monopoly status (Bogetoft and Otto, 2011, Chapter 10).

Besides a plethora of empirical applications, there has been a lively, more theoretical debate over the relative merits and downsides of both of these methods as for a long time many considered DEA and SFA as competing alternatives to each other. Although the debate may have got some more neutral tones in recent times, a huge body of work has been devoted to pinpoint the pros and cons of each method and to develop extensions of them to account for their defects (see a summary of these extensions in Fried et al., 2008). Although both methods have greatly evolved from their original forms, no clear winner of this methodological race has emerged. If anything, the comprehensive Monte Carlo simulation comparisons of the methods conducted over the years only identify different circumstances in which each method works (see e.g. Gong and Sickles, 1992; Banker et al., 1993; Andor and Hesse, 2013).

One major area of extensions for all frontier methods has been that of accounting for heterogeneity of operating environment of the firms. For example, in it is basic form DEA does not make any explicit mention of the operating environment. Instead, as Dyson et al. (2001) point out, there is an "unwritten" assumption in basic DEA that the firms should be operating in a relatively similar environment. Neither does basic SFA explicitly model the operating environment. Intuitively it is clear that comparison between any firms is meaningful only if they operate in a relatively similar environment. Otherwise, some firms may seem inefficient only because of their worse environment, not because there is some actual inefficiency present. This

concern is not new, and already about twenty years before the introduction of DEA and SFA, Hall and Winsten (1959) saw efficiency comparisons between firms operating in different environments as questionable.

Consequently, a number of solutions for accounting for the production environment heterogeneity have appeared in both DEA and SFA literature (see e.g. Coelli et al., 2005; Fried et al., 2008; section 3.6 of this introduction). One area of heterogeneity, namely *riskiness* or *uncertainty* of production environment, has not however received major attention in the frontier literature. The study of production risk, that is, the variance of output due to exogenous shocks, however has a long tradition in agricultural economics (Just & Pope, 1978; see also Moschini & Hennesy, 2001). The basic premise is, of course, that the higher variance implies riskier production environment. [3] Since production environment arguably affects output, we expect that the variation in environment has implications also for the efficiency measurement. For example, consider two farms identical in all other respects besides the weather conditions in their area. Thus we assume also that the two farmers concerned are equally efficient in turning their inputs to output(s). If farmer 1 faces more variable weather conditions, then arguably the variation in weather also affects the variation of output. As a consequence, direct efficiency measurement not acknowledging the difference in the riskiness of the operating environment confounds the shortfall in farmers' output due to weather as inefficiency.

From the previous simple example we see that examining performance without considering the variability or riskiness of that performance may mislead our analysis. Furthermore, it clearly demonstrates the importance of exogenous factors in operating environment that are mostly out of firms' control. Thus it is also important to study how these factors contribute to performance. Regarding risk, studying the variation of performance directly gives us information about the riskiness of the environment that the firm operates in. Moreover, such an analysis would give us information on how the performance variability can be controlled with the input use, for example.

[3] The traditional modelling of risk or uncertainty in production is based on typical production function with a stochastic error (see e.g. section 3.6 below). A more recent line of research, the so-called *state contingent* approach, attempts to model production uncertainty through different uncertain states of nature, which imply state-specific production (see e.g. Quiggin and Chambers, 2006 and the references therein).

Clearly, such information would be valuable for any risk-averse agent, who prefers a modest return - low risk scenario over a high return - high risk scenario, as they obviously would like to control the risk they face. In the next section we discuss in more detail how this thesis approaches the problem of risk in productivity and efficiency analysis.

2. Objectives of the thesis

Given the above background, the objectives of the thesis are broadly twofold. The first objective is methodological. It seems obvious that the chosen method should try to account for operating environment and also avoid some of the shortcomings of the so-called traditional methods. Thus this thesis systematically applies a relatively new stochastic semi-nonparametric envelopment of data (StoNED) framework developed by Kuosmanen and Kortelainen (2012). To study the effects of operating environment, we apply the notable extension of the framework by Johnson and Kuosmanen (2011). Since risk is inherently a stochastic phenomenon, the StoNED method seems suitable to study risk. As its name suggest, StoNED includes a stochastic element in its modelling framework. In this thesis, the method is applied to a wide variety of applications, ranging from aggregate productivity to energy markets. This shows the wide applicability of the chosen method, and it demonstrates the fact that heterogeneity is a concern at different levels of aggregation, be it on the level of industry or economy as a whole. The original contributions within the thesis will give a complete overview of this framework and its extensions. The use of the StoNED method and the traditional methods is also compared in the context of Finnish regulation of electricity distribution.

The second objective is naturally to examine the role of risk in productivity and efficiency analysis. Since risk or uncertainty is defined in terms of variance in this study, we utilize typical econometric tools and concepts related to heteroscedasticity to examine the issue. Very roughly, heteroscedasticity means that the variance of a certain random variable is a function of some other variables. In other words, instead of having the same variability throughout its distribution, the variance of a random variable changes due to changes in some other variables. For example, we often observe that the variance of growth rates is smaller among large firms than among smaller firms (see e.g. Hall, 1986; Dunne & Hughes, 1994). This might be because smaller firms are often younger and are yet to be so stabilized in their operations. In the context of this thesis, we naturally are interested to model the performance variation as a function of variables

which describe the operating environment of the firms. Note that in this thesis we consider production risk that is due to the variation in performance as a manifestation of risk. Obviously, uncertainty can manifest itself in production also in other ways, for example as price uncertainties in input or output prices or risks related to investments.

This thesis argues that the concept of heteroscedasticity is not yet fully understood in productivity and efficiency analysis utilizing frontier models (Saastamoinen, 2013). As discussed above, the thesis views heteroscedasticity in terms of risk. Thus we see it as an issue with economic meaning besides being purely an econometric problem. The original contributions of the thesis study this issue from different angles. More specifically, the connections between heteroscedasticity, inefficiency and risk are first studied on a conceptual level. A rather superficial gap in the literature between these topics is identified, the three concepts being closely related. The empirical applications study the heteroscedasticity issue in two different contexts. First, the connection between aggregate macro-level productivity and institutions is studied from the viewpoint of heteroscedasticity. This study suggests that the confounding relationship between corruption and aggregate productivity can be explained by the so-called macro risk effect, that we examine through heteroscedasticity. Second, heteroscedasticity is studied in the context of electricity distribution in Finland. Especially we examine how investments in underground cabling within the electricity distribution industry in Finland affect the riskiness of operations in terms of interruption costs. This is interesting from the policy perspective because both, the low general level of interruptions and their small variability can be viewed as measures of good service quality.

Before we discuss the research articles in more detail, it is necessary to define the basic concepts that are needed to understand the overall context of the thesis. Especially we need to understand the concepts of production technology and heteroscedasticity and familiarize ourselves with the methods to estimate the production technology.

3. Production technology

As much of the thesis concentrates on the empirical estimation of best practices in terms of production (output) or costs, it is critical to understand what these empirical methods estimate. For that we need to lay down the theoretical foundations of production. That is, we need to define the production possibilities of the firm and how the estimated production and cost functions relate to these possibilities. As more detailed presentations of the material of this section can be found from multiple books, the section is kept relatively brief (see e.g. Fried et al., 2008; Hackman, 2008; Coelli et al., 2005; Kumbhakar and Lovell, 2000; Färe and Primont, 1995). The notation in this section generally follows those presented in Fried et al. (2008) and Kumbhakar and Lovell (2000).

Since this thesis deals with both production and cost functions, it is important to note that the same technology which is characterized through the technical possibilities of production can also be identified through the cost minimization problem of the firm. This is known as the *duality* in economics (Diewert, 1974). Sometimes the other characterization is more suitable to model the objectives of the firm than the other. Indeed, the dual characterization allows us to model a richer set of the firms' economic objectives, not just the primal technical possibilities characterized by inputs and the corresponding outputs. For example, regulated (e.g. electricity distribution firms) companies often take their outputs as given and thus cannot be assumed to maximize production. However, it is reasonable to assume that they aim to produce their outputs with minimum costs. Thus the appropriate behavioural assumption for such companies is cost minimization as they cannot affect their revenue/profit through output adjustment (see e.g. Färe and Primont, 1995).

3.1 Firms' production possibilities

In its widest sense, we can define technology as a process where inputs **x** are transformed to outputs **y** (see e.g. Hackman, 2008). However, it is more informative to speak of *production possibility set* (PPS) when referring to

technology as this terminology explicitly defines technology as the technical possibilities of the firm. In other words, PPS consists of all combinations of inputs **x** that can produce outputs **y**. For the moment, we are speaking of multiple input, multiple output characterizations of technology and thus we refer to outputs and inputs as vectors **y** and **x**. However, when we later move from the set theoretic representation of technology to the production function representation, we consider output *y* as a scalar, which can be an aggregation of many outputs. It is nevertheless convenient to characterize the technology first explicitly for multiple outputs, a single output case being just a special case of it.

More formally, consider that we have *m* number of inputs and *s* number of outputs, which all are assumed to obtain values from the non-negative segment of the real axis. The production possibility set (PPS) can be defined as

$$T = \{(\mathbf{x}, \mathbf{y}) \in \Re_+^{m+s} \mid \mathbf{x} \text{ can produce } \mathbf{y}\} \qquad (1)$$

The definition in Equation (1) includes *all* possible input-output combinations, not only those that are observed in empirical data. Although obvious, it is often stated explicitly as an elementary assumption on the technology that all observed input-output combinations belong to the above theoretical technology. However, there are no further assumptions on this technology as yet. Nevertheless, in order to guarantee that the technology is well-behaved, we impose the following axioms listed below on the technology (see e.g. Kumbhakar and Lovell, 2000).

A1: $(\mathbf{x}, 0) \in T$ and $(0, \mathbf{y}) \notin T$.

A2: T is a closed set.

A3: For each input $\mathbf{x} \in \Re_+^m$, T is bounded.

A4: If $(\mathbf{x}, \mathbf{y}) \in T$ then $(\alpha\mathbf{x}, \mathbf{y}) \in T$ for some $\alpha \geq 1$.

A5: If $(\mathbf{x}, \mathbf{y}) \in T$ then $(\mathbf{x}, \alpha\mathbf{y}) \in T$ for some $0 \leq \alpha \leq 1$.

[A6: If $(\mathbf{x}, \mathbf{y}) \in T$ then $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in T \; \forall (-\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq (-\mathbf{x}, \mathbf{y})$]

[A7: T is a convex set.]

The assumption A1 implies that inactivity is possible and that output(s) cannot be produced without any inputs. A2 guarantees that the production

possibility set includes the technically efficient input-output combinations. In other words, the set contains its boundary since it is closed. Assumption A3 states that producing an infinite amount of output from some given amount of (finite) input is not possible. A2 and A3 together imply that the maximum amount of output that can be produced from given inputs lies on the boundary of the set. A4 and A5 impose *weak disposability* on inputs and outputs. In practice, these two assumptions allow that more inputs can used to produce the same amount of output or that less output can be produced with the same amount of inputs. The adjustment of inputs and outputs is in proportion to the factor α for all inputs or outputs when weak disposability is assumed. Assuming a more general form of adjustment, which might concern only a subset of inputs and outputs, *strong (free) disposability* should be assumed instead (Assumption A6).[4] In A6 we have assumed free disposability on both inputs and outputs. We can also assume that inputs and outputs have disposability properties that are different from each other. For efficiency measurement, disposability is essential as inefficient activities are allowed to exist by assuming disposability. Assumption A7 imposes the production possibility set to be a convex set. Assuming convex PPS is not mandatory as non-convex technologies can be assumed, but it is required to establish duality results as shall be discussed in Section 3.5. Convexity is also a critical assumption for many estimators which rely on convexity to estimate the technology.

In addition to the above assumptions, the technology is assumed to exhibit certain returns-to-scale properties. The most generic assumption is that the technology has variable returns to scale (VRS), which allows either decreasing, constant or increasing returns to scale to be present at different parts of the technology. None of the returns-to-scale assumptions need to be taken as given, as they can be tested empirically (see e.g. Banker and Natarajan, 2011). Below we start from the constant returns-to-scale (CRS)

[4] Note that free disposability implies the technology is monotonic in terms of both inputs and outputs. In other words, it means that when inputs for example increase, outputs should stay the same or increase. There are cases where for example inputs cannot be freely disposed. Weak disposability allows that increments in inputs may lead to a decrease in output, which is often labelled as input congestion (see e.g. Rødseth, 2013). Another situation where weak disposability is a reasonable assumption is in modelling bad outputs (e.g. pollution), as a bad output might not be freely disposed to keep the good output as fixed (Färe et al., 1989).

assumption and then present the assumptions of decreasing and increasing returns-to-scales in relation to that.

A8: **Constant returns to scale:** For all $\lambda > 0$, it holds that $T = \lambda T$.

The assumption A8 means that if we scale the inputs up or down by any positive factor λ, then outputs are scaled by the same factor. With this notation, increasing returns to scale means that when scaling inputs upwards with some scaling factor $\lambda_x > 1$ the increase in outputs is more than proportional, that is $\lambda_y > \lambda_x$. Decreasing returns to scale naturally means the exact opposite, that is $\lambda_y < \lambda_x$, for all $\lambda_y, \lambda_x > 1$.

Alternatively, the technology can also be represented by means of input and output sets, as shown below. Input set and output set are two equivalent ways of representing technology, and in the next section they allow us conveniently define efficiency either in terms of output expansion or input contradiction through distance functions.

Output set: The output set $P(\mathbf{x}) = \{\mathbf{y} : (\mathbf{x}, \mathbf{y}) \in T\}$ describes all possible output vectors that can be produced with a given input vector using technology *T*.

Input set: The input set $L(\mathbf{y}) = \{\mathbf{x} : (\mathbf{x}, \mathbf{y}) \in T\}$ describes all possible input vectors that can produce a given output vector using technology *T*.

Since these sets are defined in terms of the original set *T* given in Equation (1), the sets $P(\mathbf{x})$ and $L(\mathbf{y})$ inherit its corresponding properties. Thus we do not reproduce here the listing of the properties of $P(\mathbf{x})$ and $L(\mathbf{y})$ anymore (see e.g. Coelli et al., 2005). Given the sets $P(\mathbf{x})$ and $L(\mathbf{y})$, we could explicitly define also output and input isoquants and input/output efficient subsets of $P(\mathbf{x})$ and $L(\mathbf{y})$ (see e.g. Fried et al., 2008). We shall omit their exact definitions, as for our purposes it suffices to define distance functions and efficiency directly in terms of $P(\mathbf{x})$ and $L(\mathbf{y})$ only.

3.2 Distance functions

Following the definition by Koopmans (1951), improving the efficiency of production in practice means two alternative adjustments to the production process. Either you expand the level of output(s) towards the maximum output(s) that can be obtained using a given set of inputs or you contradict the amount of input(s) towards the minimum level of input(s) required to produce the given level of output(s). As we see below, distance functions characterize technology in terms of these adjustments. This implies that the Debreu-Farrell (Debreu, 1951; Farrell, 1957) measure of technical efficiency, which seeks maximal radial expansion/reduction in outputs/inputs, can be directly defined in terms of distance functions.[5]

Shepard (1953) defines the input distance function as follows:[6]

$$D_I(\mathbf{x},\mathbf{y}) = \max\{\lambda : (\mathbf{x}/\lambda) \in L(\mathbf{y})\} \qquad (2)$$

In other words, the distance function in Equation (2) seeks the maximum contradiction in inputs so that those inputs still can produce the given output level **y**. The output distance function defined by Shepard (1970), as given in Equation (3), on the other hand seeks the largest expansion in outputs so that those outputs can still be produced with the given inputs:[7]

$$D_O(\mathbf{x},\mathbf{y}) = \min\{\lambda : (\mathbf{y}/\lambda) \in P(\mathbf{x})\} \qquad (3)$$

Given the above definitions of distance functions, we see that, if either of the distance functions obtains the value of one, the adjustments of inputs or outputs are not possible. That can be regarded as efficient in the sense that no more output can be produced with given inputs or no inputs can be reduced in order to produce a given output level. It also directly follows, by construction, that $D_I(\mathbf{x},\mathbf{y}) \geq 1$ and $D_O(\mathbf{x},\mathbf{y}) \leq 1$. That is, the input amount

[5] The definition of Koopmans and the Debreu-Farrell measure do not fully coincide. Koopmans's definition implies also non-radial adjustments. This is known as a slack-problem in the Debreu-Farrell measure. Technical efficiency in Koopmans's sense is a stricter requirement than in Debreu-Farrell's sense. For our discussion. this distinction is however immaterial.

[6] See Färe and Primont (1995) for a detailed presentation of distance functions.

[7] In fact, the exact definitions use *infimum* and *supremum* instead of minimum and maximum (see e.g. Färe and Primont, 1995). For simplicity, many authors however use the more intuitive minimum and maximum definitions.

can only be reduced towards the minimum required input level and the output amount can only be expanded towards the maximum possible output. Notice that the distance functions inherit the corresponding properties of input and output sets, which obtained their properties from technology *T*.

Now that we have defined the distance functions, we can formally define the Debreu-Farrell measures of input and output technical efficiencies (*TE*) as follows:

Input (technical) efficiency:

$$\begin{aligned} &TE_I(\mathbf{x},\mathbf{y}) = \min\{\theta : \theta\mathbf{x} \in L(\mathbf{y})\} = 1 / D_I(\mathbf{x},\mathbf{y}) \\ &\Rightarrow TE_I(\mathbf{x},\mathbf{y}) \leq 1 \end{aligned} \quad (4)$$

Output (technical) efficiency:[8]

$$\begin{aligned} &TE_O(\mathbf{x},\mathbf{y}) = [\max\{\phi : \phi\mathbf{y} \in P(\mathbf{x})\}]^{-1} = [D_O(\mathbf{x},\mathbf{y})]^{-1} \\ &\Rightarrow TE_O(\mathbf{x},\mathbf{y}) \geq 1 \end{aligned} \quad (5)$$

We can see that the measurement of efficiency is intrinsically related to distance functions. This definition of technology allows us to make explicit statements about the efficiency of each firm since distance functions measure the distance of observed production to the optimal technology. In practice, we must estimate the technology (distance functions) from empirical data. Before we discuss some estimators that can be utilized in estimation, we will define technology in terms of characterizations that are more familiar to most economists, namely in terms of production and cost functions.

3.3 Production function

In the previous section, distance functions allowed us to easily characterize multi-output multi-input technology. If we can assume that the firms are producing a scalar output, either one output or some aggregate of many outputs, then technology can also be defined in terms of production function. It is soon clear that everything said about distance functions also

[8] Here it is assumed that the efficiency is measured as the ratio of optimal to the observed. Some authors define $TE_O(\mathbf{x},\mathbf{y}) = \max\{\phi : \phi\mathbf{x} \in P(\mathbf{x})\} = D_O(\mathbf{x},\mathbf{y})$ so that also $TE_O(\mathbf{x},\mathbf{y}) \leq 1$. But the definitions in (4) and (5) more naturally correspond to reduction/expansion of inputs/outputs.

applies to the production function, since production function is a special case of the above distance function characterization.

The standard definition of production function defines it as the maximal output that can be produced from given inputs (see e.g. Varian, 1992). More formally we define production function as given in Equation (6).

$$f(\mathbf{x}) = \max\{y : (\mathbf{x}, y) \in T\} \tag{6}$$

Note that only inputs are now denoted by a vector since only one output is produced. Clearly, we can immediately see that the production function defines a boundary of the technology defined in Equation (1). Of course, we have assumed that T is known and the production function can be defined as the boundary of T. Equally we could start from a known production function and define technology in terms of the known function as shown in Equation (7). This latter definition is in fact more relevant for the present purposes, as it is our aim to estimate the production function and thus recover technology through that function. Färe and Primont (1995) state that under rather mild conditions these two approaches characterize the same technology, that is $T = T'$.

$$T' = \{(\mathbf{x}, y) : f(\mathbf{x}) \geq y\} \tag{7}$$

Given the above definition of production function, we can directly define the output distance function and output technical efficiency in terms of production function as given in Equations (8) and (9). As a consequence, we can also define the production function in terms of a distance function.

$$\begin{aligned} &D_O(\mathbf{x}, y) = y / f(\mathbf{x}) \leq 1 \\ &\Rightarrow f(\mathbf{x}) = y / D_O(\mathbf{x}, y) \end{aligned} \tag{8}$$

$$TE_O(\mathbf{x}, \mathbf{y}) = [D_O(\mathbf{x}, y)]^{-1} = f(\mathbf{x}) / y \geq 1 \tag{9}$$

The equations above show the direct relation between distance functions and the production function. Again, as with distance functions, in practice we should estimate the production function from observable data in order to recover technology and assess technical efficiency.

3.4 Cost function and the duality relationship

The previous sections defined technology purely in terms of physical quantities of inputs and outputs. No assumptions about the economic behaviour of firms were made. It was only assumed that they utilize their resources as efficiently as possible. Often the success or a failure of a firm is, however, measured in economic terms. Thus it would be desirable to impose some economic behavioural assumption, such as cost minimization or revenue or profit maximization, on the firms. This way we could measure the performance of the firms with respect to these economic criteria. To achieve this, we need to identify the economic benchmarks; in other words, we need to estimate cost, revenue and profit functions. This again implies that we are trying to recover the technology of firms by using economic data besides the physical quantities of inputs and output. Besides production functions, we focus on cost functions, as both are dealt with in this thesis. But with appropriate modifications the issues of the thesis would extend to revenue and profit functions.

One concern that arises is whether the technology we identify when utilizing economic objectives differs from that where we use only physical quantities of inputs and outputs. If so, it would be difficult to say which technology is the correct one. Luckily, as already briefly stated, the duality theorem provides the means to connect the physical characterization of a technology to its economic characterization. It shows that the technology identified in either way is essentially the same as the physical production possibilities necessarily precede the economic possibilities of a firm. But before establishing duality results, we need to formally define a cost function.

With a cost function we are not restricted to examine a single output case as costs of producing multiple outputs can simply be aggregated into a single monetary value. The definitions below, of course, apply in single output case if we replace the distance functions with corresponding production function definitions. Formally, cost function can be defined as in Equation (10). The cost function is a function of input prices $\mathbf{w} \in \Re^{M}_{++}$ and outputs $\mathbf{y}$.

$$c(\mathbf{y},\mathbf{w}) = \min_{x}\{\mathbf{w}'\mathbf{x} : \mathbf{x} \in L(\mathbf{y})\} \qquad (10)$$

In other words, the cost function defines the minimum cost of producing a given level of output(s) with given input prices. Thus the choice for the firm is to choose the cost-minimizing input levels. Note that the definition of the cost function includes the conditioning $\mathbf{x} \in L(\mathbf{y})$. This gives us a clue about the duality relationship as the cost minimizing inputs naturally need to belong to the input requirement set. This implies that cost minimizing inputs have to belong to the same technology that we defined in section 3.1. Obviously, the cost minimizing inputs should be able to produce the outputs. As before, the input requirement set can be replaced with an input distance function, as in Equation (11). Thus we have defined the cost function in terms of the physical characterization of the technology. Conversely, because of duality, the input distance function can also be defined in terms of the cost function, as in Equation (12).[9]

$$c(\mathbf{y},\mathbf{w}) = \min_{x}\{\mathbf{w}'\mathbf{x} : D_I(\mathbf{x},\mathbf{y}) \geq 1\} \qquad (11)$$

$$D_I(\mathbf{x},\mathbf{y}) = \min_{w}\{\mathbf{w}'\mathbf{x} : c(\mathbf{y},\mathbf{w}) \geq 1\} \qquad (12)$$

Now it would be also straightforward to define the cost efficiency measure as a ratio of minimum costs to observed cost, as given in Equation (13). By construction, $CE(\mathbf{x},\mathbf{y},\mathbf{w}) \leq 1$.

$$CE(\mathbf{x},\mathbf{y},\mathbf{w}) = c(\mathbf{y},\mathbf{w}) / \mathbf{w}'\mathbf{x} \qquad (13)$$

Equations (11) and (12) show the duality relationship between the cost representation and the technically-based representation of technology through distance function.[10] Nevertheless, establishing the duality relationship critically depends from the axioms that we impose on the technology. Before discussing the role of axioms in establishing duality, it is, however, helpful to give also an intuitive explanation of duality. Consider a firm that is seeking to minimize its costs. It is clear that the firm cannot be at

[9] We could, of course, relate cost function directly to the production function also, as both the production function and the distance function similarly characterize the technology in physical terms. Note that we could also relate the revenue function with the output distance function/output set and the profit function with the overall technology set *T*. The latter implies that all the observed points in the overall technology set have costs equal or greater than the minimum cost, that is: $T = \{(\mathbf{x},\mathbf{y}) : \mathbf{w}'\mathbf{x} \geq c(\mathbf{y},\mathbf{x})\}$.

[10] See proof in Färe and Primont (1995).

the cost minimizing point if it could reduce the amount of one input and still produce the same amount of output. In that sense, the physical measure of technical efficiency must be related to the cost-based measurement of efficiency and the technology that they characterize should be the same. Thus technical efficiency is necessary but not yet sufficient condition for overall cost efficiency. There is also an allocative efficiency part in overall cost efficiency. This means that, although the firm is obtaining a maximum output from given inputs, with the given input prices the firm is using the inputs in wrong proportions. We, however, shall omit the discussion of allocative efficiency here.

3.5 Axioms of production and duality

In section 3.1, we presented the axioms that a technology should satisfy in order to guarantee that the technology is well-behaved and physically feasible. In this section, we further discuss the practical meaning of these axioms with the help of some intuitive examples. We especially relate the convexity axiom to the duality theorem. Moreover, we highlight that the axioms impose necessary structure on technology so that we are able to estimate the technology with the methods introduced in the next section.

Free disposability is a relatively intuitive assumption to make. You can expect that two workers can dig a 10-meter long trench in an hour if one worker can do it. In this example, the output is kept fixed, but the input is increased. On the other hand, we could turn the situation around by saying that if a worker can dig a 10-meter trench in an hour, he can also dig a 5-meter one in the same time. As already stated, the importance of disposability lies in the fact that it allows inefficient actions to be present.

Consider next the convexity of input sets, which directly follows from the convexity of T. This implies that a convex combination of input vectors in the input set should also belong to the set. Assume for example that one piece of machinery (K) can replace two workers (L). If we have two input vectors $(L,K)_i = [(4,0)_1;(0,2)_2]$ which are both capable of producing a certain level of output, then we would expect that an input vector $(2,1)$ also can achieve that. The new input vector corresponds to a convex combination of the original technologies, where both technologies get the weight of 0.5.

Convexity thus implicitly assumes that inputs are continuously divisible. If divisibility is not a reasonable assumption, then some non-convex estimation technology can be considered (see e.g. Keshvari and Kuosmanen, 2013).

Convexity matters for production functions also. A production function is often assumed to be (quasi)concave. This follows from the convexity assumption for the input set $L(\mathbf{y})$. It means that a convex combination of input vectors that both belong to the PPS restricted above by $f(\mathbf{x})$, belong to the PPS also. But more importantly, the concavity assumption implies that the production function exhibits diminishing returns to inputs. This is one of the most fundamental laws in production economics. Now compare this to the cost function. Cost function is a convex function of outputs if *T* is convex. [11] This property is directly analogous (or rather, consequence of it) to the concavity property of production function. Indeed, analogously to Equation (10), the cost function can be defined in terms of the production function as $c(\mathbf{y},\mathbf{w}) = \min_x\{\mathbf{w}'\mathbf{x} : y \le f(\mathbf{x})\}$. The convexity of the cost function in outputs implies non-decreasing marginal costs for inputs, which can be considered as the cost equivalent of diminishing returns of inputs on the production side. What can be seen here is that convexity is crucial in order to establish duality.

The above discussion shows that axioms do have practical meaning in terms of basic economic fundamentals. In addition, axioms impose some structure on production technology, allowing its meaningful estimation. The set in Equation (1) is too general to be estimated without any further assumptions about the technology, as it only characterizes the feasible input-output pairs without giving any guidance about the underlying structure behind these input-output correspondences. In some methods, we already impose a lot of structure on the technology by directly assuming some

[11] With a slight digression, it is good to note at this point that, for natural monopolies, the cost function needs to be sub-additive in the sense that $C(y_1 + y_2) \le C(y_1) + C(y_2)$. This means that the cost of producing outputs y_1 and y_2 separately in two firms is higher than in one firm (Baumol, 1977). Schmalensee (1978) shows that a necessary condition for the existence of natural monopoly in distribution/transmission type industries is that the cost function is concave in output. If a function is concave, it is also sub-additive. As will be discussed later, it is problematic if functional forms which are convex in outputs are used to estimate costs in cases where the characteristics of production clearly imply natural monopoly, as is the case in electricity distribution.

functional form for the production, cost or distance function beforehand. Such an approach would be commonly referred to as parametric. It is then afterwards tested whether the estimated parameters of the function acceptably satisfy the axioms. Unfortunately, some functional forms violate by construction some of the assumptions. Alternatively, some methods rely only on the axioms themselves, without assuming any specific functional form. Such an approach would be nonparametric. In the next section, we will deal with both of these approaches to estimate the technology in question.

3.6 Estimation of technology

In previous sections, we have defined technology either in terms of distance, production or cost functions. Regardless of how we define it, in practice we need to estimate it from the observed empirical data. In economics, there has been a long econometric tradition of production function estimation at least since the work of Cobb and Douglas (1928). The development of production function estimation was intimately related to the development of productivity measurement (Griliches, 1996). However, as mentioned in Section 1, the conventional economic theory and the corresponding econometric approach have both assumed that technical inefficiency has been resolved (see e.g. Kumbhakar and Lovell, 2000). The only source of deviation from the production function is assumed to be purely due to random statistical noise. When we explicitly introduce inefficiency to the model, the traditional (econometric) estimators do not apply anymore since we introduce another source of deviation from the optimum.

In this section, we briefly discuss the three estimation frameworks, namely the DEA, SFA, and StoNED estimators, that explicitly acknowledge the presence of technical inefficiency. We keep the discussion of the methods in this section very brief and concentrate only on their most significant differences. This is because the detailed accounts of each method are presented in multiple books and later on in this thesis (see e.g. the first research article of this thesis; Fried et al., 2008; Coelli et al., 2005, Cooper et al. 2000, Kumbhakar and Lovell, 2000). Thus repeating that discussion here is not worthwhile.

Our discussion of the estimators is most easily done in the context of a general production model. This helps us to compare how the estimators differ with respect to the general model. The model is given in Equation (14), where y is the observed output, $f(\mathbf{x})$ is the production function which is a function of inputs $\mathbf{x}$, $\boldsymbol{\beta}$ is the parameter vector to be estimated and $\boldsymbol{\varepsilon}$ is an error term representing the deviation of the observed production from the estimated one. For the moment, we shall not make any specific assumptions about the error term.

$$y = f(\mathbf{x};\boldsymbol{\beta}) + \boldsymbol{\varepsilon} \qquad (14)$$

Few notes are in place. First, to simplify the discussion, we shall consider only the production function in this section. The discussion would naturally extend to the estimation of cost and distance functions also. Second, in order to more closely relate the general production model presented here to the subsequent contributions of the thesis, we assume that only a single output is produced within the general model. This is because the basic StoNED method utilized in the thesis allows only one output to be present.

3.6.1 Stochastic frontier analysis (SFA)

We start with the stochastic frontier analysis (SFA) due to its close relationship with the traditional estimation framework. SFA was introduced almost simultaneously by Aigner et al. (1977) and Meeusen and van den Broeck (1977). As does the traditional econometric estimation of production functions, also SFA assumes a specific functional form for $f(\mathbf{x})$, for example Cobb-Douglas, translog, or constant elasticity of substation (CES) form. However, in contrast to the traditional framework, stochastic frontier approach assumes that $\boldsymbol{\varepsilon}$ is composed of two parts, as the equation $\boldsymbol{\varepsilon} = \mathbf{v} - \mathbf{u}$ shows. The vector $\mathbf{v}$ is the usual stochastic noise that you would have in any regression model. This would constitute the only source of deviation in the traditional framework. The vector $\mathbf{u}$ describes the shortfall in output due to inefficiency. Thus in SFA, firms are inefficient with respect to the stochastic frontier $f(\mathbf{x}) + \mathbf{v}$.

As usual in regression analysis, it is assumed that **v** is distributed symmetrically with zero mean. Normal distribution is practically always assumed for **v**. For **u**, a one-sided distribution is assumed, as inefficiency can only reduce output. Examples of this are half-normal, exponential and gamma distribution, from which the first is by far the most typical assumption. It is also usually assumed that **v** and **u** are independent from each other and from the inputs $\mathbf{x}$. This composed form of error has significant implications for the estimation, and it is a major departure from the traditional framework which assumes only symmetrically distributed **v** to exist. In SFA, $\boldsymbol{\varepsilon}$ cannot have a symmetric and zero mean distribution as the composed error is a convolution of the symmetric and non-symmetric part. This convolution is problematic as our target of interest, inefficiency, is convoluted with the often uninteresting part, namely noise (see e.g. Amsler, Lee, and Schmidt, 2009). Since we want to separate these two, an alternative estimation method which specifically accounts for this characteristic of the composed error is needed.[12] In this thesis, the noise part is also of a certain interest to us since we are examining production risk.

In SFA literature, there are two main approaches to estimate the parameters of the production function subject to the composed error. The first approach is based on maximum likelihood. Given some distributional assumptions on both **v** and **u**, the log-likelihood function can be formulated in terms of $\boldsymbol{\varepsilon}$. Assuming for example the standard normal – half-normal assumptions $\mathbf{v} \sim N(0, \sigma_v^2)$ and $\mathbf{u} \sim N^+(0, \sigma_u^2)$, the log-likelihood for sample of $i = 1, ..., N$ firms is as shown in Equation (15).

$$\ln L(y \mid \boldsymbol{\beta}, \sigma, \lambda) = \text{constant} - N \ln \sigma + \sum_{i=1}^{N} \ln \Phi\left(-\frac{\varepsilon_i \lambda}{\sigma}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \varepsilon_i^2 \qquad (15)$$

[12] An early related discussion can be found in Aigner and Chu (1968) and in Førsund and Jansen (1977). These authors set so-called *average practice* functions against *best practice/frontier* functions. In this terminology, the traditional model in Equation (14) with symmetrically distributed errors with zero mean could equally well describe the average practice of firms, as the estimated function goes through the middle of the cloud of points. But again, if we wish the follow the stance of conventional economic theory without inefficiency, then such model can be considered to describe the best practice of firms subject only to statistical noise.

where $\sigma^2 = \sigma_u^2 + \sigma_v^2$ and $\lambda = \sigma_u / \sigma_v$. This likelihood function is maximized in order to find the estimates for the parameters $(\boldsymbol{\beta}, \sigma, \lambda)$. It is relatively straightforward to see from Equation (15) that if $\lambda = 0$ (or equivalently $\sigma_u = 0$) the model collapses to the standard maximum likelihood formulation of the ordinary least squares (OLS) regression problem, assuming normally distributed errors. With $\sigma_u = 0$, there is no inefficiency present, as $\sigma_u = 0 \Rightarrow u = 0$, and consequently the overall error is $\boldsymbol{\varepsilon} = \mathbf{v}$. For example, Kuosmanen and Fosgerau (2009) suggested testing the appropriateness of stochastic frontier specification from the skewness of residuals $\hat{\varepsilon}$. By construction, given the distributional assumptions, the overall error $\boldsymbol{\varepsilon} = \mathbf{v} - \mathbf{u}$ should be negatively skewed if inefficiency is present.

An alternative approach is based on OLS estimation. It has been long known in frontier literature that all parameters of the production frontier, expect the intercept, can be consistently estimated with OLS (see e.g. Olson, Schmidt, and Waldman, 1980; Greene, 1980). Thus the estimation can be broken down into two parts. First, we estimate the other parameters with OLS and, in the second step, correct the intercept so that the estimated function corresponds to a frontier. Basically this means that the estimated function is shifted upwards. This approach is generally referred to as the method of moments (MoM) or the Modified OLS approach since the second step correction is based on the moments of the OLS residuals of the first stage.[13] More formally, to estimate the first step with OLS, the original model needs to be reformulated so that the error has a zero mean. This can be done in the following fashion. We have taken out the intercept β_0 from the parameter vector $\boldsymbol{\beta}$ to explicitly show the bias in the intercept.

$$\begin{aligned} y &= \left[\beta_0 - E(\mathbf{u})\right] + f(\mathbf{x};\boldsymbol{\beta}) + \mathbf{v} - \left[\mathbf{u} - E(\mathbf{u})\right] \\ &= \left[\beta_0 - E(\mathbf{u})\right] + f(\mathbf{x};\boldsymbol{\beta}) + \boldsymbol{\varepsilon}^* \end{aligned} \qquad (16)$$

[13] We shall not digress here to the realm of parametric deterministic methods, namely the *parametric* or *goal programming* approach proposed by Aigner and Chu (1968) and the *corrected OLS* (COLS) proposed originally by Winsten (1957). Besides being deterministic, these methods have their own limitations discussed for example by Kumbhakar and Lovell (2000) and Florens and Simar (2005).

Since $\mathbf{v}$ is assumed to have zero mean, the expectation of $\boldsymbol{\varepsilon}^* = \mathbf{v} - [\mathbf{u} - E(\mathbf{u})]$ is also zero as shown below.

$$E(\boldsymbol{\varepsilon}^*) = E[\mathbf{v} - [\mathbf{u} - E(\mathbf{u})]] = E(\mathbf{v}) - E(\mathbf{u}) + E(\mathbf{u}) = 0$$

Thus the original intercept is biased by the amount of expected inefficiency. Notice that no specific distributional assumptions about **v** or **u** have been made at this stage as OLS does not explicitly require any such assumptions. In the second step, after (16) has been estimated with OLS, we make the usual normal and half-normal distributional assumptions about **v** and **u**. Given these assumptions, the theoretical second and the third central moments (m_2 & m_3) of the composite error term can be written as in (17) and (18).

$$m_2 = \left[\frac{\pi - 2}{\pi}\right]\sigma_u^2 + \sigma_v^2 \qquad (17)$$

$$m_3 = \left(\sqrt{\frac{2}{\pi}}\right)\left[1 - \frac{4}{\pi}\right]\sigma_u^3 \qquad (18)$$

Equating the above theoretical moments with their sample counterparts $\hat{m}_2$ and $\hat{m}_3$, which can be estimated using the residuals $\hat{\boldsymbol{\varepsilon}}^*$, we can easily solve the formulas for the variance parameters, as shown in Equations (19) and (20). Note that the values $\hat{\boldsymbol{\varepsilon}}^*$ can be used to estimate the variance and the skewness of the original $\hat{\boldsymbol{\varepsilon}}$ as both variance and skewness are invariant to a constant location change in the distribution, which the shift with $E(u)$ ultimately is.

$$\hat{\sigma}_u^2 = \left(\frac{\hat{m}_3}{\sqrt{2/\pi}(1 - 4/\pi)}\right)^{2/3} \qquad (19)$$

$$\hat{\sigma}_v^2 = \hat{m}_2 - (1 - 2/\pi)\hat{\sigma}_u^2 \qquad (20)$$

Since we have assumed that *u* has a half-normal distribution, it follows that $E(\mathbf{u}) = \sqrt{\frac{2}{\pi}}\hat{\sigma}_u$ and thus the intercept can be estimated as

$$\hat{\beta}_{0;MOLS} = OLS \text{ intercept} + \sqrt{\frac{2}{\pi}}\hat{\sigma}_u .$$

After the frontier has been estimated (either with ML or MOLS), the estimation of the inefficiency term **u** often follows. In the normal-half-normal model, conditional mean $E(u_i \mid \varepsilon_i)$ derived by Jondrow et al. (1982) is the most typical point estimate of u_i. Confidence intervals for the efficiency estimates can also be obtained (Horrace & Schmidt, 1996). [14]

Accounting for production environment within SFA models is relatively straightforward. Most of the typical approaches rely on parameterizing the parameters (mean and/or variance) of the inefficiency distribution as a function of operating environment variables, which are commonly referred to as z-variables in the literature (see the survey by Kumbhakar & Lovell, 2000). It is assumed that the z-variables are not part of the production technology as such but that they affect the (in)efficiency of producers. For example, the model by Kumbhakar, Ghosh and McGuckin (1991) parameterizes the distribution of inefficiency, as given in (21).

$$u_i \sim \left| N(\mathbf{z}_i'\boldsymbol{\gamma}, \sigma_u^2) \right| \qquad (21)$$

This model defines the inefficiency as positive truncation of a normal distribution such that the mean of the un-truncated distribution can differ between observations. The second research article of the thesis discusses these types of models in more detail.

3.6.2 Data envelopment analysis (DEA)

In this section, we very briefly cover another widely applied estimation method to estimate production frontiers subject to inefficiency. We went to some mathematical details with SFA as some of its features are intrinsically related to the StoNED framework. The same applies to DEA, but for ease of exposition, we avoid presenting any mathematical details of DEA. As opposed to SFA, the basic premise of DEA is that it easily incorporates multiple outputs into it. This can be achieved also in SFA when the cost

[14] The performance of maximum likelihood and MoM estimators has been compared for example by Olson et al. (1980) and Coelli (1995). The performance of each depends on the sample size and the relative contribution of inefficiency compared to noise. Generally ML is more efficient but MOLS is somewhat more robust to distributional assumptions, as they are avoided in the first step.

function or distance functions are parametrically estimated, but the basic SFA production frontier models always assume a single output. DEA, of course, can also be used to estimate cost, revenue, profit, and distance functions, but again we keep the context of the general production model out of the discussion.

The motivation for DEA arose from the challenge of measuring productivity and efficiency in a multi-output multi-input context. Using a simple index of outputs over inputs is generally challenging as it is not obvious how to aggregate and weight different outputs and inputs. DEA was developed as an estimation method to obtain optimal weights for outputs and inputs. Already Farrell (1957) introduced the basic idea of DEA, the details of which were further formalized by Afriat (1972). Farrell proposed to measure the efficiency of firms with respect to a surface that envelopes all the observations. However, only two decades later this approach was branded as data envelopment analysis. Charnes, Cooper, and Rhodes (1978; abbreviated commonly as CCR) operationalised the insights of Farrell and Afriat as a simple linear programming problem and popularized the application of DEA for wider audience both within practitioners and academics.

Purely as a mathematical problem, DEA finds the frontier of observations such that the efficiency of each firm is maximized and is 100% at the maximum. In other words, DEA attempts to find the tightest possible envelopment of the data such that the efficiency of each firm is maximized. This is called the minimum extrapolation principle within the DEA literature. The envelopment is piecewise linear, and the shape of this envelopment is fully dictated by the economic theory and the available data by assuming the axioms of convexity and free disposability. Note that no specific functional form assumption is made concerning the function $f(\mathbf{x})$ in Equation (14). The shape constraints only restrict the estimated function to follow some regularity conditions, but they do not assume any specific form for the function. In addition to convexity and free disposability, some assumption about the returns to scale must be made. The CCR-model assumes constant returns to scale (CRS), whereas variable returns to scale (VRS) extension of DEA was suggested by Banker et al. (1984). All of the assumptions can be relaxed in turn, and for example relaxing the convexity

assumption leads to a free disposal hull (FDH) estimator proposed by Deprins et al. (1984). Moreover, given the application, the model is formulated either in output or input orientation, depending whether we consider adjustments in outputs or inputs, respectively. In some applications, firms' outputs are seen fixed, and thus efficiency can only be improved by adjusting the input use. For example, the previously mentioned electricity distribution case would fall to this category as energy demand (energy delivered), the number of customers, and the network size cannot be adjusted much by the companies.

Besides the functional form, the second major difference between DEA and SFA is their assumption about the error term $\boldsymbol{\varepsilon}$ in Equation (14). Since DEA aims to envelope all data perfectly, it implicitly assumes that all deviations from the frontier are due to inefficiency, i.e. $\boldsymbol{\varepsilon} = \mathbf{u}$. No statistical noise is allowed in basic DEA. Moreover, the DEA frontier is fully dictated by the outermost observations, which by construction are 100% efficient. Indeed, both SFA and DEA estimate relative performance measures. But conceptually they differ in what they assume about the best performers. In SFA, a firm is practically never estimated to be exactly 100% efficient because the continuous distributions assumed for $\mathbf{v}$ and $\mathbf{u}$ imply that the probability of a single point being exactly on the frontier is zero.

Lastly, we briefly cover some methods to show how to take operation environment into account in DEA. Following the categorization by Coelli et al. (2005), the ways to account for production environment in DEA setting can be roughly divided at least into three approaches.

The first alternative is to include z-variables as non-discretionary (non-controllable) variables into the linear programming problem directly. This is generally referred to as the one-stage approach (see Syrjänen, 2003, for a detailed discussion and references). This approach basically restricts the benchmark set for the firms to make them more comparable in terms of these environmental variables. Often the restriction would be placed so that a firm cannot be compared to those with a better environment. The main limitation of this approach is that the direction of the effects of environmental variables needs to be known beforehand.

Secondly, one might assume that different DEA frontiers should be estimated for firms operating in different environments. This approach is

generally referred to as the frontier separation approach (Charnes, et al., 1981). With this approach, it is possible to compare the best possible performances in different environments to each other if the different frontiers are compared with respect to some overall frontier. The problem is to know beforehand how to divide the firms into their respective subgroups in order to estimate the separate frontiers. This is not always unambiguous as there might be many possible ways to divide the sample.

The third option is to obtain the usual DEA efficiency scores at first. In the second stage, a regression of the obtained efficiencies on environmental variables can be conducted to study how environmental variables affect the efficiencies. Then either the efficiency scores or the original outputs/inputs can be adjusted with these effects and, in the latter case, a DEA model would be rerun with the adjusted variables (see e.g. Fried et al., 1999, 2002). The advantage of this approach is that it gives detailed information about the effects of z-variables in the second step, and we do not need to assume anything about the effects of z-variables a priori. However, as recognized in the literature, a direct implementation of a simple regression of efficiency scores on z-variables is not advisable (see e.g. Simar and Wilson, 2007; Banker and Natarajan, 2008; Johnson and Kuosmanen, 2012).

3.6.3 Stochastic semi-nonparametric envelopment of data (StoNED)

The last estimator we introduce is the StoNED estimator. Since the full details of StoNED framework are given in the first research article of the thesis, this section only outlines the relation of StoNED to SFA and DEA in terms of the general production model given in Equation (14).

In general, the StoNED estimator attempts to combine the best features of the traditional SFA and DEA estimators. As opposed to DEA, StoNED incorporates statistical noise into its framework. This is desirable at least for three reasons. First, the main interest of this thesis, risk, is a phenomenon that is inherently stochastic. Second, if we do not allow for noise, we implicitly assume that our data is measured without any error and no specification error of the model exists. These often are too strong assumptions to be made. Third, the stochastic noise term gives the estimator

a more statistical/econometric grounding. This is desirable especially for statistical inference. In a mathematical programming based approach without noise, such as DEA, it is not directly obvious how statistical inference should be conducted.[15]

On the other hand, unlike SFA, StoNED does not make any functional form assumptions regarding $f(\mathbf{x})$. Similarly to DEA, it bases its estimation of technology on some general axioms about the technology. This is a desirable property of both StoNED and DEA as it often is difficult to justify any specific functional form over another. For example, the Cobb-Douglas form assumes perfect substitutability between inputs. Another unfortunate feature of the Cobb-Douglas form is that it does not properly model the economies of scope since it favours output specialization over joint-production. This is problematic in modelling for example the cost efficiency of electricity distribution firms, where the typical outputs of distribution firms are necessarily jointly produced (Kuosmanen et al., 2013). Some more flexible functional forms may solve some of these problems, but they often violate convexity and monotonicity. Since the StoNED method combines these features of SFA and DEA, it is a more general estimation framework. In fact, as Kuosmanen and Johnson (2010) show, DEA can be formulated as a special case of the StoNED framework. The same applies to SFA also.

In practice, the StoNED estimation procedure has many similarities to the SFA approach presented above. In the MOLS framework, StoNED replaces the parametric OLS in the first step with a nonparametric counterpart, namely with Convex Nonparametric Least Squares (CNLS, Hildreth, 1954; see also Kuosmanen, 2008). Otherwise the procedure is exactly the same. An alternative to the method of moments is the pseudo-likelihood approach formulated in terms of CNLS residuals. Both approaches need some parametric distributional assumptions for inefficiency and noise in order to separate them. But the first CNLS stage is fully non-parametric. Thus it is appropriate to call StoNED a semi-nonparametric method.

Taking account of operating environment is rather straightforward. Johnson and Kuosmanen (2011) extend the typical StoNED model and

[15] Simar and Wilson (2000) suggest a bootstrap-based inference for nonparametric efficiency measures such as DEA.

include z-variables into the first-stage CNLS estimation. This one-stage estimator is preferred over a two-stage estimator where the estimation of z-effects is left to the second step. Omitting z-variables in the first stage may cause the two-step estimator to be biased due to the omitted variable/endogeneity problem (Wang & Schmidt, 2002; Schmidt, 2010).[16] We discuss this z-variable extension of StoNED in more detail in the first research article of the thesis.

[16] See also Johnson and Kuosmanen (2012), who compare the performance of two-step and one-step estimators of z-effects.

4. Heterogeneity, Heteroscedasticity and Risk

In this section, we briefly introduce the three concepts that are essential to this thesis. We start by examining why heterogeneity in general is important in performance measurement. We then introduce one specific kind of heterogeneity, namely heteroscedasticity. We highlight how some basic econometric tools dealing with heteroscedasticity directly lend themselves to the study of production risk. Lastly, we discuss the sufficiency of variance as a risk measure. The usability of econometric tools of heteroscedasticity to study risk critically depends on whether we consider variance as an adequate measure of risk.

4.1 Heterogeneity

In economics, the term heterogeneity is often reserved to mean a deviation from the representative agent assumption. That is, the acting agents (e.g. consumers or firms) are not identical in this case. Within this thesis, we however extend the term to mean also the heterogeneity of operating environment. These two are often indistinguishable from each other as economic agents adapt their behaviour according to their environment, which can, consequently, change due to this behaviour (see e.g. Kirman, 2006). From the point view of performance evaluation and efficiency measurement, the critical problem is that often heterogeneity confounds our measures of performance (Greene, 2004). To further clarify this, let us consider few examples.

First, in the economics of growth, it is widely acknowledged that institutions play an important role in economic development (Hall & Jones, 1999). Institutions such as political centralization, property rights, labour market laws and cultural or societal norms undoubtedly are heterogeneous among countries. Certainly these factors also contribute to the ability of countries to utilize their resources efficiently (see. e.g. Moroney and Lovell, 1997; Adkins, Moomaw, and Savvides, 2002 for some early contributions in frontier literature). Thus a direct assessment of the performance only in terms of GDP, capital and labour seems inadequate as it neglects the effect of institutions on the resource utilization capabilities of a nation. It is only after

we have extracted out the effect of institutions when we can start to compare nations in their efficiency of transforming labour and capital into gross domestic product.

Second, consider agricultural production. Arguably the heterogeneity is present in the analysis of agricultural production as agricultural producers are faced with great spatial differences for example in soil quality and weather (see e.g. Just, 2003). Variation in these may again manifest itself as variation in crops. The challenge is that these variations in the operating environment may be misinterpreted as efficiency differences (Greene, 2004; O'Donnell et al., 2010). In some sense, different environments imply slightly different technologies for each producer. But even if we explicitly exclude the problem of different technologies, the heterogeneity of operation environment still poses a challenge. For example, in electricity distribution, firms' technology can be assumed to be relatively similar due to some technical norms that the firms need to meet. Nevertheless, two companies which both distribute electricity through similar cables may operate in highly different environments in terms weather and forest density, for example.

In light of the above examples, we see that any measurement of productivity or efficiency without acknowledging heterogeneity is likely to confound inefficiencies with other sources of variation in productivity. Setting a common standard seems unacceptable if firms operate under highly different conditions. Indeed, it is often these factors outside the firms' control that are the underlying reason of performance differentials between firms. This is not to downplay the importance of managerial or technical inefficiency. But their importance might be overstated if we do not account for other sources of performance variation also.

One concern is that what aspects of operation environment or producer-specific heterogeneity are relevant enough to be taken account. Here probably the only thing that can be done is to rely on the expertise of the researcher to know what to include into the model. But, for the researcher, the problem is that not all relevant aspects are necessarily observed. When introducing frontier methods in Section 3, we indeed assumed that operation environment can be represented as a function of some observable variables. If however some relevant aspects of the environment are unobserved, it probably would be desirable to extend the methods to take this into account.

While there have been developments on how SFA methods explicitly account for some unobserved factors, similar developments are still waiting in the StoNED setting (see e.g. Kopsakangas-Savolainen and Svento, 2011, for discussion in SFA context). Although this is an obvious target for future development, we do not extend our examination into the realm of unobserved heterogeneity. This is because we are especially interested in the sources of heteroscedasticity in this thesis. This obviously requires us to model heteroscedasticity as a function of some observable variables. Moreover, let's recall that StoNED by construction is a flexible method in terms of technology, and thus it is likely to capture some of the unobserved firm-specific heterogeneity already in the technology parameters.

4.2 Heteroscedasticity

Next we briefly introduce one specific form of heterogeneity, namely heteroscedasticity. We do this in the context of the basic linear regression model. Since heteroscedasticity is routinely dealt in any econometrics textbook, we keep our examination very brief (for detailed presentations see e.g. Verbeek, 2008; Greene, 2008).

Consider a general regression model for a sample of N observations given in Equation (22).

$$y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{22}$$

where

y is the $N \times 1$ vector of a dependent variable.

$\mathbf{X}$ is the $N \times K$ matrix of K independent variables, including a column of one for intercept.

$\boldsymbol{\beta}$ is the $K \times 1$ vector of parameters to be estimated.

$\boldsymbol{\varepsilon}$ is the $N \times 1$ vector of error terms.

The ordinary least squares OLS estimator of the parameters $\boldsymbol{\beta}$ would be $\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. The standard Gauss-Markov assumptions for the OLS estimator assume that the error term is homoscedastic. This means that the variance of the error term is constant across all observations N.

Mathematically, $Var[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \sigma^2 \mathbf{I}$, where σ^2 is the unknown error variance and $\mathbf{I}$ is an $N \times N$ identity matrix. Thus the diagonal elements of the variance-covariance matrix are the same for every observation. Heteroscedasticity is present if $Var[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \sigma^2 \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is some matrix with elements $[\psi_1, \psi_2, ..., \psi_N]$ on its diagonal. For simplicity, we assume that the off-diagonal elements are zero, implying that no autocorrelation of the errors is present. Under heteroscedasticity, the OLS estimator is still unbiased and consistent. It is, however, possible to obtain a more efficient estimator. More importantly, standard statistical inference is not valid under heteroscedasticity, as the usual t- and F-statistic are invalid. This can be dealt with by using heteroscedasticity-robust standard errors (White, 1980). Furthermore, tests to detect heteroscedasticity have been suggested for example by Breusch and Pagan (1980) and White (1980).

Under heteroscedasticity, generalized least squares (GLS) estimators can be more efficient than OLS. GLS weights the observations in terms of their variation so that the observations with highest variance are typically given the lowest weight. Thus GLS requires that the weighting matrix $\boldsymbol{\Psi}$ is known beforehand. For the moment, we assume that this matrix is known and that we have a very general form of heteroscedasticity such that $\sigma_i^2 = \sigma^2 \psi_i$**.** Then the GLS estimator for β is $\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{y}$. Effectively our data is weighted with the weights $1/\psi_i$ since heteroscedasticity is proportional to the weights ψ_i.

In practice, the matrix $\boldsymbol{\Psi}$ has to be estimated and thus a feasible generalized least squares (FGLS) estimation procedure must be applied. To estimate the matrix $\boldsymbol{\Psi}$, a functional form for heteroscedasticity has to be assumed. Often *multiplicative* heteroscedasticity, as shown in Equation (23), is assumed (Harvey, 1976; see also Verbeek, 2008, p. 96). It guarantees that we obtain positive estimates of variances. Note that the better efficiency properties of FGLS compared to OLS hinge on knowing the correct weights. If we assume a wrong form of heteroscedasticity in FGLS, it is not guaranteed that with small sample sizes FGLS would outperform OLS. Generally, FGLS is however justified in asymptotic sense at least.

$$\sigma_i^2 = \sigma^2 \exp(\mathbf{z}_i \boldsymbol{\alpha}) \quad (23)$$

In Equation (23), $\mathbf{z}_i$ is usually the vector of the original independent variables of the model or some subset of them and $\boldsymbol{\alpha}$ is the corresponding parameter vector which tells the effect of these variables on error variance. In SFA context, however, z-variables are often assumed to be some other variables than the original input variables.

In modelling heteroscedasticity, another common strategy is to assume that the error variance is a linear function of some variables, as in Equation (24). Using this formulation in FGLS is problematic as the linear form does not guarantee that the variances are non-negative. The form however suffices to test heteroscedasticity and to examine the heteroscedasticity effects of z-variables.

$$\sigma_i^2 = \mathbf{z}_i \boldsymbol{\alpha} \quad (24)$$

Regardless of the functional form assumed, the estimation of the parameters $\boldsymbol{\alpha}$ is usually based on the OLS residuals of the model in Equation (22), although direct maximum likelihood estimation in a single step is also possible. Note after the initial estimates of the weights have been obtained, FGLS can be reapplied and the whole procedure can be iterated a number of times to further improve the estimates of $\boldsymbol{\beta}$ (Greene, 2008).

Heteroscedasticity in the SFA context has received some attention as both the parameters of the production technology and the efficiency measures can be biased if heteroscedasticity in the inefficiency term is not accounted for (Caudill and Ford, 1993; Caudill et al., 1995), that is, if we wrongly assume that $\sigma_{i,u}^2 = \sigma_u^2 \quad \forall i = 1,...,N$. Heteroscedasticity can also be in the stochastic noise term *v*, as noted by Hadri (1999), but its consequences are less severe (see e.g. Kumbhakar & Lovell, 2000). The SFA models that attempt to account for heteroscedasticity in the inefficiency term parameterize the standard deviation or the variance of the inefficiency distribution as a function of z-variables (see e.g. Alvarez et al., 2006). In the DEA literature, there has been less concern about heteroscedasticity, partly because of the nonparametric nature of DEA. We postpone any detailed

discussion of heteroscedasticity in the StoNED context to the first research article of the thesis.

Simulation studies do not provide a clear-cut picture of the effects of heteroscedasticity on the performance of the estimators presented in Section 3.6 (Banker et al. 2004; Kuosmanen & Kortelainen, 2012; Andor and Hesse, 2013). Overall, the magnitude and even the direction of this effect are affected by the following issues: presence or absence of noise, the relative importance of inefficiency and noise in the data generating process, whether we consider heteroscedasticity in inefficiency or noise or in both, whether we examine the estimation of a frontier or the point estimate of inefficiency, and sample size.

4.3 Variance and risk measurement

It is quite straightforward to see that the methods outlined in the previous section easily adapt themselves to the study of production risk and its causes if we assume that variance is a sufficient measure of risk. Indeed, the variance of the error ε in (22) directly translates to variation in output also. However, by assuming variance as the appropriate measure of risk, we often make some implicit assumptions about the nature of risk. Nevertheless, it is not directly obvious whether these assumptions suit to all risky situations. Before that discussion, it is however good to make a distinction between uncertainty and risk as these concepts are often confounded with each other.

Traditionally, a risky situation has been characterized such that the acting agent is able to assign some probabilities to the possible future events in such a situation. In the case of uncertainty, this assignment is commonly not possible (Knight, 1921; Chavas, 2004). As Knight (1921) defines it, risk is measurable whereas uncertainty is not. Chavas (2004), however, points out that these definitions depend much on how we define probabilities and their possible existence. First, the ease of assigning probabilities varies from one situation to another. It is difficult to assign any probabilities for rare events such as plane crashes, whereas it is easy to derive the probabilities of a dice throw. The former case falls under uncertainty, whereas a throw of a dice is risky. It is also easy to elicit probabilities in a case of repeated events. A producer for example may have a rather good understanding of the

probability of a malfunction in a production line, as the producer observes the functioning of the line over a long period. If the production line operates under relatively unchangeable and controllable conditions, the frequency of malfunctions is a good measure of their probability. But production may also be subject to changeable conditions which are hard to measure. As the probability is likely to differ under these varying conditions, it is difficult to form an objective assessment about the probability of an event. The second issue is that the probabilities in many cases are subjective. Consequently, there may not be any agreement about the 'correct' probabilities. But, arguably, human beings at least implicitly assign some probabilities on future events regardless of how *uncertain* the events are. Following Chavas, as long as the future outcome of an event is not known beforehand and some probabilities or likelihoods (subjective or objective) are assigned to the outcomes, it is only of secondary importance whether we call this situation *risky* or *uncertain*. Furthermore, as it was pointed out already by Arrow (1951), the unmeasurable Knightian uncertainties often lead us to the same conclusions as explicit probabilities. Thus we use these terms interchangeably within this thesis.

The use of variance as a measure of risk is very intuitive. If we are offered two bets, A and B, which have the same expected value but with the difference that B has a higher change for higher losses and wins, then arguably we view the bet B as the riskier one. Often the formal origin for the use of variance as a risk measure is attributed to the financial portfolio theory of Markowitz (1952). Basically, the Markowitz's (1952) mean-variance model brought the minimization of variance as another objective for the investor next to the maximization of the expected return. Subsequently, multiple treatments on the limitations of the mean-variance setup have been written by Markowitz (1959) and for example by Hanoch & Levy (1969). A more detailed discussion and a list of references to alternative approaches can be found for example in Grootveld & Hallerbach (1999) and Antle (2010).

The most typical objections to variance as a risk measure concern the fact that variance is a symmetric measure and it does not take the skewness of the distributions properly into account. For example, if we consider the variability of crop yields in agricultural production only in terms of variance, we implicitly assume that very low and very high yields are equally likely. In

some gambles, symmetrically distributed returns/losses may be a reasonable assumption. But for crop yields it is often argued that observing extraordinary good crops should be less likely than mean – and less than mean crops – since extraordinary crops require ideal weather conditions, which rarely occur (see e.g. Gallagher, 1987). One of the more popular alternatives has been so-called down-side risk models. These models emphasize the importance of losses over the possible returns for investor, and thus they put more weight (or all of it) on the left hand side of the return distribution.

However, considering the focus of this thesis, we restrict ourselves to the traditional variance-based measurement of risk, for two reasons. Firstly, the concept of heteroscedasticity is the main unifying thread of the thesis. Since models of heteroscedasticity are models of variance, it is natural to restrict our attention to this measure of risk. Moreover, the development of frontier models that deal with production uncertainty is still relatively limited and is focused on heteroscedasticity.[17] For example, O'Donnell et al. (2010) lists only few papers that explicitly discuss risk in the context of technical efficiency estimation. Thus it seems warranted to investigate the relatively underexploited models of heteroscedasticity in more detail before extending the research agenda to more novel models of risk, with possibly skewed risk distributions. Secondly, and maybe more importantly, skewness is reserved to represent the presence of inefficiency in frontier models, not the presence of production risk. Skewness-based measurement of risk in the frontier context could be challenging, due to convolution of risk and inefficiency. As will be discussed in the third research article of the thesis, too similar distributional assumptions on inefficiency and noise (risk) make it impossible to distinguish them from the overall error (Amsler, Lee, and Schmidt, 2009). Thus, we restrict ourselves to the symmetric measures of risk to more clearly differentiate it from skewed inefficiency.

[17] See e.g. Kumbhakar (2002), Wang (2002), and Bera & Sharma (1999).

5. Summary of the research papers

In this section, we briefly cover the contributions of each research article that are included in the thesis. The thesis consists of five research articles. All of them are connected, either through the topic of heteroscedasticity or through a methodological choice. The logic for the ordering of the articles is the following. The first two articles concentrate more on introducing the StoNED method and comparing it with the traditional frontier estimators. As opposed to Section 3.6 of this introduction, the comparison is done in an empirical setting, namely within the context of Finnish electricity distribution regulation. The last three articles then concentrate on the topic of heteroscedasticity and production risk. Article 3 reviews literature on production risk and heteroscedastic SFA models. Articles 4 and 5 then deals with two empirical applications related to heteroscedasticity, still applying the StoNED method.

5.1 Research article 1

Stochastic nonparametric approach to efficiency analysis: A Unified Framework

This handbook chapter outlines the StoNED framework in a detailed manner and acts as methodological review for the thesis. Many of the topics in the chapter go beyond the topics dealt with in this thesis. However, the chapter further motivates the reader to see the additional benefits of the StoNED estimator when compared with the traditional frontier estimators. There is a particular emphasis on that the traditional approaches can be seen as special cases of the more general StoNED framework. The chapter also includes a detailed examination of heteroscedasticity in the StoNED context. This examination shows that the many well-known approaches dealing with heteroscedasticity in econometrics are relatively straightforwardly applicable in the StoNED context also.

5.2 Research article 2

What is the best practice for benchmark regulation of electricity distribution? Comparison of DEA, SFA and StoNED methods

This research article is the result of the project that the authors did for the Finnish electricity market regulator *Energiamarkkinavirasto* (EMV).[18] In 2010, EMV decided to refine the tools that it uses to measure the cost efficiency of electricity distribution companies. EMV adopted the StoNED method as their preferred tool for the regulatory period of 2012-2015. In their selection criteria for the method, EMV emphasized that the method should be flexible in production technology and account for stochastic noise and heterogeneity in production environment. Previously, EMV used the average of DEA and SFA efficiency scores to determine the cost efficiency of companies. This approach supposedly mitigated the potential problems arising from using only one method. The paper argues that such an approach is statistically unsound and proposes StoNED to be used instead. The performance of the StoNED method was compared to the traditional estimators with an empirical comparison of efficiency scores and economic outcomes of the regulation. A simulation study was also conducted to compare the methods under a fixed data generation setting.

This research article serves as an important illustration about the practical uses of efficiency estimation methods. The article also continues the methodological discussion started in the book chapter (see previous section) about the differences between traditional methods and the StoNED framework. The issue of heteroscedasticity is not explicitly discussed here but the importance of production environment heterogeneity is highlighted throughout the paper as the electricity distribution companies in Finland operate in highly varying geographical and climatic conditions.

5.3 Research article 3

Heteroscedasticity or Production Risk? A Synthetic View

This review article compares two branches of literature, namely the literature on production risk and that on heteroscedastic stochastic frontier models. To

[18] Kuosmanen et al. (2010).

our knowledge, no such systematic comparison of the two branches of literature had been conducted before this article, although the fields have a number of similarities. The purpose of the article is to establish connections between the production risk and efficiency literature. Thus, thematically the article ties together the two main topics of the thesis. The link between the fields is built by utilizing the concept of heteroscedasticity. It is argued that although heteroscedastic stochastic frontier models do often neglect risk considerations, they can be interpreted to model production risk. Lastly, some of the challenges of simultaneous treatment of risk and efficiency are also discussed.

5.4 Research article 4

Is Corruption Grease, Grit, or a Gamble? Corruption Increases Variance of Productivity Across Countries

This research article is the author's first exploration of the topic of heteroscedasticity and heterogeneity within frontier methods. Subsequently, it dictated the thematic focus for the rest of the thesis. Initially, the interest was just to study the effect of corruption on productivity. However, during the research process, the authors observed that the effects of corruption seemed to be more related to the variability of productivity than to the level of productivity. As a way to understand this phenomenon, heteroscedasticity was considered as an intuitive way to model this relationship.

The general view in the literature has been that corruption hinders economic development. This argument is generally referred to as "sand in the wheels" hypothesis. However, some real-life observations have induced some researchers to suggest that under certain circumstances corruption in fact might be beneficial for the economic performance of countries, acting as "grease" in the wheels of economic development. This paper considers a completely alternative view. It reckons that rather than a direct determinant of economic performance corruption should be considered as a macro risk. In other words, we interpret corruption as a gamble, since it seems to increase the variability of productivity among countries. This hypothesis is considered to be more general than the traditional views because it allows that, with relatively high levels of corruption, a country can achieve either high or low

productivity levels. The empirical examination indeed shows that corruption has significant effect on the variability of productivity.

5.5 Research article 5

Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland

This research article is a partial continuation of the paper introduced in section 5.2. It returns to the same electricity distribution application. However, instead of focusing on cost efficiency, the article deals with the quality of service of the distribution companies. This is a topical issue as there are pressures in Finland to develop the distribution system towards underground cabling based system.

The article examines the quality of service of electricity distribution in terms of interruption costs and their variability. The paper considers that, besides the low level of interruption costs, the low variability of interruption costs is also a sign of good service quality. Underground cabling is the most significant investment target to affect interruptions. The article shows that underground cabling expectedly decreases the level of interruption costs. However, we also observe that underground cabling does not have significant decreasing effect on the variability of interruption costs. In some instances the effect might even be increasing. This effect we explain by the fact that the costs of interruptions are significantly higher for companies with higher underground cabling levels. Such companies often operate in areas of high population density, and, subsequently, the costs of interruption are likely to be high as a large number of households are affected by the interruption. Interruption costs of underground cabling are increased also because underground cables are more costly to install and repair than air cables.

The article also compares two alternative ways to set the quality targets in quality regulation. The current practice is based on the average of the companies' own past performance. However, average is a very volatile measure as it can be greatly affected by single years with a high number of interruptions. Moreover, usage of average does not incentivize to improve upon a poor previous performance. The article suggests that the targets should be set in terms of the best observed operations. For this purpose, we

estimate an interruption cost frontier with StoNED and refer to it as the *quality frontier*. Comparison of the obtained target values shows that the targets produced by the quality frontier are significantly more stable than the targets produced by the average of own performance.

6. Concluding remarks

Here we briefly summarize the main contribution of the thesis. Clearly, heteroscedasticity has economic and practical interpretations beyond being just an econometric problem. In the empirical applications of the thesis, we have interpreted heteroscedasticity mainly in terms of risk. We have shown that this risk may realize itself in the aggregate productivity or quality of service type of contexts. We have also noted that there exist some gaps in the literature, namely between the traditional models of production risk and heteroscedastic stochastic frontier models. Methodological choice has been shown to be important in practical applications of the frontier methods. Indeed, important regulatory outcomes may crucially depend on the chosen efficiency estimation method. These contributions correspond to the objectives of the thesis set in Section 2. Of course, the views presented in this thesis should not be viewed as definite resolutions to the topic. But if one is to examine the risk-efficiency nexus in future, the issues covered in this thesis seem an unavoidable starting point for that research.

References

Abramowitz, M. (1956). Resource and output trends in the United States since 1870. *American Economic Review*, 46 (2): 5-23.

Adkins, L.C., Moomaw, R.L., and Savvides, A. (2002). Institutions, freedom, and technical efficiency. *Southern Economic Journal*, 69 (1), 92-108.

Afriat, S. (1972). Efficiency estimation of production functions. *International Economic Review*, 13: 568-598.

Aigner, D. J. and Chu, S. F. (1968). On Estimating the Industry Production Function. *American Economic Review*, 58 (4): 826-839.

Aigner, D. J., Lovell, C. A. K., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6 (1): 21-37.

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

Amsler, C., Lee, Y. H., and Schmidt, P. (2009). A survey of stochastic frontier models and likely future development. *Seoul Journal of Economics*, 22 (1), 6-27.

Andor, M. and Hesse, F. (2013). The StoNED age: the departure into a new era of efficiency analysis? A Monte Carlo comparison of StoNED and the ''oldies'' (SFA and DEA). *Journal of Productivity Analysis*, DOI 10.1007/s11123-013-0354-y.

Antle, J. M. (2010). Asymmetry, partial moments, and production risk. *American Journal of Agricultural Economics*, 92 (5): 1294-1309.

Arrow, K. J. (1951). Alternative Approaches to the Theory of Choice in Risk-Taking Situations. *Econometrica*, 19 (4), 404-437.

Banker, R. D., Charnes, A., and Cooper W. W. (1984). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Management Science*, 30 (9): 1078-1092.

Banker, R. D., Gadh, V. M., and Gorr, W. L. (1993). A Monte Carlo comparison of two production frontier estimation methods: corrected ordinary least squares and data envelopment analysis. *European Journal of Operational Research*, 67, 332-343.

Banker, R. D., Chang, H., and Cooper, W. W. (2004). A simulation study of DEA and parametric frontier models in the presence of heteroscedasticity. *European Journal of Operational Research*, 153: 624-640.

Banker, R. D. and Natarajan, R. (2008). Evaluating contextual variables affecting productivity using data envelopment analysis. *Operations Research*, 56 (1): 48-58.

Banker, R. D. and Natarajan, R. (2011). Statistical tests based on DEA efficiency scores. Chapter 11 in: Cooper, W.W., Seiford, L., and Zhu, J. (Eds.), *Handbook on Data Envelopment Analysis*, 2nd edition. Springer, pp. 273-295.

Baumol, W. J. (1977). On the proper cost tests for natural monopoly in a multiproduct industry. *American Economic Review*, 67 (5): 809-822.

Bera, A. K., and Sharma, S. C. (1999). Estimating production uncertainty in stochastic frontier production function models. *Journal of Productivity Analysis*, 12 (3): 187-210.

Bogetoft, P. and Otto, L. (2011). *Benchmarking with DEA, SFA, and R*. International Series in Operations Research and Management Science, 157. Springer.

Breusch, T. and Pagan, A. (1980). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47: 1287-1294.

Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2 (6): 429-444.

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

Caudill, S. and Ford, J. (1993). Biases in frontier estimation due to heteroscedasticity. *Economics Letters*, 41(1): 17-20.

Caudill, S., Ford, J., and Gropper, D. (1995). Frontier estimation and firm-specific inefficiency measures in the presence of heteroscedasticity. *Journal of Business and Economic Statistics*, 13 (1): 105-111.

Chavas, J.P. (2004). *Risk Analysis in Theory and Practice*. Elsevier Academic Press, San Diego.

Cobb, S. and Douglas, P. (1928). A theory of production. *American Economic Review*, 18, 139-165.

Coelli, T. (1995). Estimators and hypothesis test for a stochastic frontier function: a Monte Carlo analysis. *Journal of Productivity Analysis*, 6: 247-268.

Coelli, T. J., Rao, D. S. P., O'Donnell, C. J., and Battese, G. E. (2005). *An Introduction to Efficiency and Productivity Analysis*. New York, USA, Springer.

Coelli, T. and Lawrence, D. (eds.) (2006). *Performance Measurement and Regulation of Network Utilities*. United Kingdom, Edward Elgar.

Cooper, W. W., Seiford, L. M., and Tone, K. (2000). *Data Envelopment Analysis: a comprehensive text with models, applications, references, and DEA-Solver software*. Massachusetts, USA, Kluwer Academic Publishers.

Debreu, G. (1951). The coefficient of resource utilization. *Econometrica*, 19 (3): 273-292.

Deprins, D., Simar, L., and Tulkens, H. (1984). Measuring Labor Inefficiency in Post Offices, in Marchand, M., Pestieau, P. and Tulkens, H. (eds.), *The Performance of Public Enterprises: Concepts and Measurements*, Elsevier Science Publisher, B.W., pp. 243-267.

Diewert, W. E. (1974). Applications of duality theory. In: Intriligator, M. D. and Kendrick, D. A. (Eds.), *Frontiers of Quantitative Economics*, Volume II, North-Holland Publishing Company.

Dunne, P. and Hughes, A. (1994). Age, size, growth and survival: UK companies in the 1980s. *The Journal of Industrial Economics*, 42 (2): 115-140.

Dyson, R. G., Allen, R., Camacho, A. S., Podinovksi, V. V., Sarrico, C. S., and Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research*, 132: 245-259.

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

Florens, J-P. and Simar, L. (2005). Parametric approximations of non-parametric frontiers. *Journal of Econometrics*, 124: 91-116.

Førsund, F. R. and Jansen, E. S. (1977). On estimating average and best practice homothetic production functions via cost functions. *International Economic Review*, 18 (2): 463-476.

Fried, H.O., Schmidt, S.S., and Yaisawarng S. (1999). Incorporating the operational environment into a nonparametric measure of technical efficiency. *Journal of Productivity Analysis*, 12: 249-267.

Fried, H.O., Lovell, C.A.K., Schmidt, S.S., and Yaisawarng S. (2002). Accounting for environmental effects and statistical noise in data envelopment analysis. *Journal of Productivity Analysis*, 17: 157-174.

Fried, H. O., Lovell C. A. K., and Schmidt S. S., eds. (2008). *The Measurement of Productive Efficiency and Productivity Growth*. New York, USA, Oxford University Press Inc.

Färe, R., Grosskopf, S., Lovell, C. A. K., and Pasurka, C. et al. (1989). Productivity Comparisons When Some Outputs are Undesirable: A Nonparametric Approach. *The Review of Economics and Statistics*, 71 (1): 90-98.

Färe, R. and Primont, D. (1995). *Multi-Output Production and Duality: Theory and Applications*. Kluwer Academic Press, USA.

Gallagher, P. (1987). U.S. soybean yields: Estimation and forecasting with nonsymmetric disturbances. *American Journal of Agricultural Economics*, 69 (4): 796-803.

Gong, B-H. and Sickles, R. C. (1992). Finite sample evidence on the performance of stochastic frontiers and data envelopment analysis using panel data. *Journal of Econometrics*, 51: 259-284.

Greene, W. H. (1980). Maximum likelihood estimation of econometric frontier production model. *Journal of Econometrics*, 13, 27-56.

Greene, W. H. (2004). Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organizations's panel data on national health care systems. *Health Economics*, 13: 959-980.

Greene, W. H. (2008). *Econometric Analysis*, 6th edn. Pearson Prentice Hall, New Jersey, USA.

Griliches, Z. (1996). The discovery of the residual: A historical note. *Journal of Economic Literature*, 34: 1324-1330.

Grootveld, H. and Hallerbach, W. (1999). Variance vs. downside risks: Is there really that much difference? *European Journal of Operational Research*, 114: 304-319.

Hackman, S. T. (2008). *Production economics: Integrating the microeconomic and engineering perspectives*. Springer.

Hadri, K. (1999). Estimation of a doubly heteroscedastic stochastic frontier cost function. *Journal of Business and Economic Statistics*, 17 (3): 359-363.

Hall, B. H. (1986). The relationship between firm size and firm growth in the U.S. manufacturing center. *NBER Working Paper Series*, WP. No. 1965.

Hall, M. and Winsten C. (1959). The ambiguous notion of efficiency. *Economic Journal*, 69 (273): 71-86.

Hall, R.E. and Jones, C.I. (1999). Why do some countries produce so much more output per worker than others? *The Quarterly Journal of Economics*, 114 (1): 83-116.

Hanoch, G. and Levy, H. (1969). Efficiency analysis of choices involving risk. *The Review of Economic Studies*, 36 (3): 335-346.

Hildreth, C. (1954). Point estimates of ordinates of concave functions. *Journal of the American Statistical Association*, 49 (267): 598-619.

Harvey, A. C. (1976). Regression models with multiplicative heteroscedasticity. *Econometrica*, 44 (3): 461-465.

Horrace, W. C. and Schmidt, P. (1996). Confidence statements for efficiency estimates from stochastic frontier models. *Journal of Productivity Analysis*, 7: 257-282.

Johnson, A. and Kuosmanen, T. (2011). One-stage estimation of the effects of operational conditions and practices on productive performance: asymptotically normal and efficient, root-n consistent StoNEZD method. *Journal of Productivity Analysis*, 36 (2): 219-230.

Johnson, A. and Kuosmanen, T. (2012). One-stage and two-stage DEA estimation of the effects of contextual variables. *European Journal of Operational Research*, 220: 559-570.

Jondrow, J., Lovell, C., Materov, I., and Schmidt, P. (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics*, 19 (2-3): 233-238.

Just, R. E. (2003). Risk research in agricultural economics: opportunities and challenges for the next twenty-five years. *Agricultural Systems*, 75: 123-159.

Just, R. E. and Pope, R. D. (1978). Stochastic specification of production functions and economic implications. *Journal of Econometrics*, 7 (1): 67-86.

Keshvari, A. and Kuosmanen, T. (2013). Stochastic non-convex envelopment of data: Applying isotonic regression to frontier estimation. *European Journal of Operational Research*, 231: 481-491.

Kirman, A. (2006). Heterogeneity in economics. *Journal of Economic Interaction and Coordination*, 1: 89-117.

Knight, F. (1921). *Risk, Uncertainty and Profit*. Boston: Houghton Mifflin.

Koopmans, T.C. (1951). An analysis of production as an efficient combination of activities. In T.C. Koopmans (ed.) *Activity Analysis of Production and Allocation*. Cowles Commission for Research Monograph No. 13. New York: John Wiley and Sons.

Kopsakangas-Savolainen, M. and Svento, R. (2011). Observed and unobserved heterogeneity in stochastic frontier models: An application to the electricity distribution industry. *Energy Economics*, 33: 304-310.

Kumbhakar, S. C. (2002). Specification and estimation of production risk, risk preferences and technical efficiency. *American Journal of Agricultural Economics*, 84 (1): 8-22.

Kumbhakar, S. C., Ghosh, S., and McGuckin, J. (1991). A generalized production frontier approach for estimating determinants of inefficiency in U.S. dairy farms. *Journal of Business and Economic Statistics*, 9(3): 279-286.

Kumbhakar, S. C. and Lovell, C. A. K. (2000). *Stochastic Frontier Analysis*. New York, USA, Cambridge University Press.

Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. *Econometrics Journal*, 11: 308-325.

Kuosmanen, T. and Fosgerau, M. (2009). Neoclassical versus frontier production models? testing for the skewness of regression residuals. *Scandinavian Journal of Economics*, 111 (2): 351-367.

Kuosmanen, T. and Johnson, A. (2010). Data envelopment analysis as nonparametric least squares regression. *Operations Research*, 58 (1): 149-160.

Kuosmanen, T., Kortelainen, M., Kultti, K., Pursiainen, H., Saastamoinen, A., and Sipiläinen, T. (2010). Sähköverkkotoiminnan kustannustehokkuuden estimointi StoNED-menetelmällä: ehdotus tehostamistavoitteiden ja kohtuullisten kustannusten arviointiperusteiden kehittämiseksi kolmannella valvontajaksolla 2012–2015 (in Finnish). Available from: www.emvi.fi (accessed 6.1.2014).

Kuosmanen, T. and Kortelainen, M. (2012). Stochastic non-smooth envelopment of data: Semi-parametric frontier estimation subject to shape constraints. *Journal of Productivity Analysis*, 38 (1): 11-28.

Kuosmanen, T., Saastamoinen, A., and Sipiläinen, T. (2013). What is the best practice for benchmark regulation of electricity distribution? Comparison of DEA, SFA and StoNED methods. *Energy Policy*, 61: 740-750.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7: 77-91.

Markowitz, H. (1959). *Portfolio selection: efficient diversification of investments*. Cowles Foundation Monograph #16, Wiley, New York.

Mas-Colell, A., Whinston, M. D., and Green, J. R. (1995). *Microeconomic Theory*. Oxford University Press, New York, USA.

Meeusen, W. and van den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, 18 (2): 435-445.

Moroney, J. and Lovell, C.A.K. (1997). The relative efficiencies of market and planned economies. *Southern Economic Journal*, 633 (4): 1084-1093.

Moschini, G., and Hennessy D.A. (2001). Uncertainty, risk aversion, and risk management for agricultural producers. In Gardner, B.L. and Rausser, G.C. (eds.) *Handbook in Agricultural Economics*, Volume 1, Part A. Downloadable at http://www.sciencedirect.com/ (Accessed on 12.11.2013).

O'Donnell, C. J., Chambers, R. G., and Quiggin, J. (2010). Efficiency analysis in the presence of uncertainty. *Journal of Productivity Analysis*, 33 (1): 1-17.

Olson, J. A., Schmidt, P., and Waldman, D. M. (1980). A Monte Carlo study of stochastic frontier production functions. *Journal of Econometrics*, 13: 67-82.

Quiggin, J. and Chambers, R. G. (2006). The state-contingent approach to production under uncertainty. *The Australian Journal of Agricultural and Resource Economics*, 50: 153–169

Rødseth, K. L. (2013). A note on input congestion. *Economics Letters*, 120: 599-602.

Saastamoinen, A. (2013). Heteroscedasticity or production risk? A synthetic view. *Journal of Economic Surveys*, DOI: 10.1111/joes.12054.

Schmalensee, R. (1978). A note on economies of scale and natural monopoly in the distribution of public utility services. *The Bell Journal of Economics*, 9 (1): 270-276.

Schmidt, P. (2010). One-step and two-step estimation in SFA models. *Journal of Productivity Analysis*, 36 (2): 201-203.

Shepard, R. W. (1953). *Cost and Production Functions*. Princeton, Princeton University Press.

Shepard, R. W. (1970). *The Theory of Cost and Production Functions*. Princeton, Princeton University Press.

Simar, L. and Wilson P. W. (2000). Statistical inference in nonparametric frontier models: The state of the art. *Journal of Productivity Analysis*, 13: 49-78.

Simar, L. and Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics*, 136: 31-64.

Syrjänen, M. (2003). Data envelopment analysis in planning and heterogeneous environments. Doctoral dissertations, Helsinki School of Economics.

Syverson, C. (2011). What determines productivity? *Journal of Economic Literature*, 49 (2): 326-365.

Wang, H. (2002). Heteroscedasticity and non-monotonic efficiency effects of a stochastic frontier model. *Journal of Productivity Analysis*, 18 (3): 241-253.

Wang, H. and Schmidt, P. (2002). One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. *Journal of Productivity Analysis*, 18 (2): 129-144.

Varian, H. (1992). *Microeconomic Analysis*. New York: W.W. Norton & Company Inc.

Verbeek, M. (2008). *A Guide to Modern Econometrics*. John Wiley & Sons, 3rd ed., Chichester, England.

White, H. (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica*, 48: 817-838.

Winsten, C. B. (1957). Discussion of Mr. Farrell's paper. *Journal of Royal Statistical Society, Series A, General*, 120 (3): 282-284.

Part II: Original research articles

Article 1

Timo Kuosmanen; Andrew Johnson; Antti Saastamoinen. Stochastic nonparametric approach to efficiency analysis: A Unified Framework. Unpublished manuscript. Forthcoming in J. Zhu (Ed.) *Handbook on DEA*, Springer.

Stochastic nonparametric approach to efficiency analysis:

A Unified Framework

Timo Kuosmanen[1], Andrew Johnson[2,1], Antti Saastamoinen[1]

1) School of Business, Aalto University, 00100 Helsinki, Finland.

2) Department of Industrial and Systems Engineering, Texas A&M University, TX 77840, USA.

1. Introduction

Efficiency analysis is an essential and extensive research area that provides answers to such important questions as: Who are the best performing firms and can we learn something from their behavior? [1] What are the sources of efficiency differences across firms? Can efficiency be improved by government policy or better managerial practices? Are there benefits to increasing the scale of operations? These are examples of important questions we hope to resolve with efficiency analyses.

Efficiency analysis is an interdisciplinary field that spans such disciplines as economics, econometrics,[2] operations research and management science,[3] and engineering, among others. The methods of efficiency analysis are utilized in several fields of application including agriculture, banking, education, environment, health care, energy, manufacturing, transportation, and utilities, among many others. Efficiency analysis is performed at various different scales. Micro level applications range from individual persons, teams, production plants and facilities to company level and industry level efficiency assessments. Macro level applications range from comparative efficiency assessments of production systems or industries across countries to efficiency assessment of national economies. Indeed, efficiency improvement is one of the key components of productivity growth (e.g., Färe et al., 1994), which in turn is the primary driver of economic welfare. The benefits to understanding

[1] We will henceforth use the term "firm" referring to any production unit that transforms inputs to output, including both non-profit and for-profit organizations. The firm can refer to an establishment (facility) or sub-division of a company or to an aggregate entity such as an industry, a region, or a country.

[2] Observe that 13 of the 100 most cited articles published in a leading field journal, the *Journal of Econometrics*, are efficiency analysis papers, including Simar and Wilson (2007) that has 436 citations, making it the #32 most cited paper in the journal in just 6 years from its publication (citations data gathered from Scopus, Nov 25, 2013).

[3] In operations research and management science, Charnes et al. (1978) ranks #1 as most cited article published in the *European Journal of Operational Research* (EJOR) and Banker et al. (1984) is the #1 most cited article in *Management Science*, two of the leading journals of this field (the flagship journals of EURO and INFORMS, respectively). In fact, Charnes et al. article has more than 5 times more citations than the 2nd most cited paper in EJOR (Nov 25, 2013).

the relationship between efficiency and productivity and quantifying efficiency cannot be overstated. In words of Paul Krugman (1992, p. 9), "*Productivity isn't everything, but in the long run it is almost everything. A country's ability to improve its standard of living over time depends almost entirely on its ability to raise its output per worker*." Note that macro-level performance of a country is an aggregate of the individual firms operating within that country. Therefore, sound micro-foundations of efficiency analysis are critical for the integrity of productivity and efficiency analysis at macro level.

Unfortunately, there currently is no commonly accepted methodology of efficiency analysis, but the field is divided between two competing approaches: Data envelopment Analysis (DEA) and Stochastic Frontier Analysis (SFA). [4]

Data envelopment analysis (DEA, Farrell, 1957; Charnes et al., 1978) is an axiomatic, mathematical programming approach to efficiency analysis. DEA's main advantage compared to econometric, regression-based tools is its nonparametric treatment of the frontier, building upon axioms of production theory such as free disposability (monotonicity), convexity (concavity), and constant returns to scale (homogeneity). DEA does not assume any particular functional form for the frontier or the distribution of inefficiency. It's direct, data-driven approach is helpful for communicating the results of efficiency analysis to decision-makers. However, the main shortcoming of DEA is that it attributes all deviations from the frontier to inefficiency. This is often a heroic assumption.

Stochastic frontier analysis (SFA, Aigner, Lovell, Schmidt, 1977; Meeusen and Vanden Broeck, 1977) is often, incorrectly, viewed as a direct competitor of DEA. The key strength of SFA is its probabilistic modeling of deviations from the frontier, which are decomposed into a non-negative inefficiency term and an idiosyncratic error term that accounts for omitted factors such as unobserved heterogeneity of firms and their operating environments, random errors of measurement and data processing, specification errors, and other sources of noise. In contrast to DEA, SFA utilizes parametric regression techniques, which require *ex ante* specifications of the functional forms of the frontier and the inefficiency distribution. Since the economic theory rarely justifies a particular functional form, flexible functional forms such as translog are frequently used. However flexible functional forms often violate axioms of production theory, whereas imposing the axioms will reduce flexibility. In summary, the DEA and SFA methods are not direct competitors but rather complements: in the tradeoff between DEA and SFA something is sacrificed for something to be gained. Namely DEA does not model noise, but is able to impose axiomatic properties and estimate the frontier non-parametrically, while SFA cannot impose axiomatic properties, but has the benefit of modeling inefficiency and noise.

[4] Citation statistics of some of the key papers provide undisputable evidence about the significant influence of this field. The four most cited papers are Charnes et al. (1978) with 6,152 citations, Banker et al. (1984) with 3,415 citations, Farrell (1957) with 3,296 citations, and Aigner et al. (1977) with 1,875 citations (Scopus, Nov 25, 2013).

Bridging the gap between axiomatic DEA and stochastic SFA was for a long time one of the most vexing problems in the field of efficiency analysis. The recent works on convex nonparametric least squares (CNLS) by Kuosmanen (2008), Kuosmanen and Johnson (2010), and Kuosmanen and Kortelainen (2012) have led to the full integration of DEA and SFA into a unified framework of productivity analysis, which we refer to as *stochastic nonparametric envelopment of data* (StoNED).[5]

We see the development of StoNED as a paradigm shift for efficiency analysis. It is no longer necessary to decide if modeling noise is more important than imposing axioms of production theory: we can do both using StoNED. The unified framework of StoNED offers deeper insights to the foundations of DEA and SFA, but it also provides a more general and flexible platform for efficiency analysis and related themes such as frontier estimation and production analysis. Further, a number of extensions to the original DEA and SFA methods have been developed over the past decades. The unified StoNED framework allows us to combine the existing tools of efficiency analysis in novel ways across the DEA-SFA spectrum, facilitating new opportunities for further methodological development.

The main objective of this chapter is to provide an updated and elaborated presentation of the CNLS and StoNED methods, the most promising new tools for axiomatic nonparametric frontier estimation and efficiency analysis under stochastic noise. Our secondary objective is to extend the scope of the StoNED method in several dimensions. This chapter provides the first extension of the StoNED method to the general case of multiple inputs and multiple outputs. We also consider quantile estimation using StoNED, and present a detailed discussion of how to model heteroscedasticity in the inefficiency and noise terms.

The rest of this chapter is organized as follows. Section 2 introduces the unified StoNED framework and its special cases by reviewing alternative sets of assumptions that motivate different estimation methods applied in productivity analysis. Our focus is explicitly on the axiomatic DEA-style approaches. Section 3 presents the CNLS regression as a quadratic programming problem. Section 4 discusses the intimate connections between CNLS and DEA, and introduces a step-wise C^2NLS estimator. Section 5 further develops the step-wise estimation approach for the StoNED estimator. Section 6 reviews some important extensions to the StoNED, including the multiplicative formulation (Section 6.1), observations from multiple time periods that make up a panel data (Section 6.2), directional distance functions (DDF) for modeling multiple output variables (Section 6.3), and quantile regression formulation (Section 6.4). The model of contextual variables that represent operational conditions or practices is examined in detail in Section 7.

[5] The term StoNED was coined by Kuosmanen (2006). By request of referees, Kuosmanen and Kortelainen (2012) used the term stochastic "non-smooth" envelopment, as their model specification involves parametric distributional assumptions. In this chapter we show that the distributional assumptions can be relaxed: see Sections 5.2.3 and 6.2.

Testing of heteroscedasticity and modeling heteroscedasticity of inefficiency and noise using a doubly-heteroscedastic model discussed in Section 8. Finally, Section 9 concludes with discussion of some promising avenues of future research.

2. Unified frontier model

To maintain direct contact with the SFA literature, we introduce the unified model of frontier production function in the multiple input, single output case. Multiple outputs can be modeled using cost functions (see Kortelainen and Kuosmanen, 2012, Section 4.4; and Kuosmanen, 2012) and distance functions. A general multi-input multi-output directional distance function model will be introduced in Section 6.3.

Production technology is represented by a frontier *production function* $f(\mathbf{x})$, where $\mathbf{x}$ is a m-dimensional input vector. [6] Frontier $f(\mathbf{x})$ indicates the maximum output that can be produced with inputs $\mathbf{x}$, and hence the function $f(\mathbf{x})$ characterizes the boundary of the production possibility set. We assume that function f belongs to the class of continuous, monotonic increasing, and globally concave functions that can be non-differentiable (we denote this class as F_2). This is equivalent to stating that the production possibility set satisfies the classic DEA assumptions of free disposability and convexity. In contrast to SFA, no specific functional form for f is assumed.

The observed output y_i of firm i may differ from $f(\mathbf{x}_i)$ due to inefficiency and noise. We follow the SFA literature and introduce a composite error term $\varepsilon_i = v_i - u_i$, which consists of the inefficiency term $u_i > 0$ and the stochastic noise term v_i, formally,

$$\begin{aligned} y_i &= f(\mathbf{x}_i) + \varepsilon_i \\ &= f(\mathbf{x}_i) - u_i + v_i, \quad i = 1,...,n \end{aligned} \tag{1}$$

Variables u_i and v_i ($i = 1,...,n$) are random variables that are assumed to be statistically independent of each other as well as of inputs $\mathbf{x}_i$. We assume that the inefficiency term has a positive mean and a constant finite variance, that is, $E(u_i) = \mu > 0$ and $Var(u_i) = \sigma_u^2 < \infty$. We further assume zero mean noise with a constant finite variance, that is, $E(v_i) = 0$ and $Var(v_i) = \sigma_v^2 < \infty$. Assuming σ_u^2 and σ_v^2 are constant across firms is referred to as homoscedasticity; models with heteroskedastic inefficiency and noise will be discussed in Section 8. For the sake of generality and to maintain the fully nonparametric orientation, we do not introduce any distributional assumptions for u_i or v_i at this point.

[6] For clarity, we denote vectors by bold lower case letters (e.g., $\mathbf{x}$) and matrices by bold capital letters (e.g., $\mathbf{Z}$). All vectors are column vectors, unless otherwise indicated. Note: $\mathbf{x}'$ denotes the transpose of vector $\mathbf{x}$.

However, some estimation techniques to be introduced below require additional parametric assumptions.

In model (1), the deterministic part (i.e., production function *f*) is defined analogous to the DEA literature, while the stochastic part (i.e., composite error term ε_i) is defined similar to SFA. As a result, model (1) encompasses the classic models of the SFA and DEA literature as its constrained special cases. Note that in this chapter we use the term "model" in the sense of the econometric literature to refer to the description of the data generating process (DGP). DEA and SFA are alternative estimators or methods for estimating the production function *f*, the expected inefficiency μ, and the firm-specific realizations of the random inefficiency term u_i. We note that in the DEA literature it is common to use the term "model" for the linear programming problem (e.g., LP model) or other mathematical programming formulations for computing the estimator. To avoid confusion, we will follow the econometric terminology and refer to equation (1) and the related assumptions as the model, whereas DEA, SFA, CNLS, and StoNED are referred to as estimators. In this terminology, "DEA model" or "SFA model" refer to the specific assumptions regarding the variables of model (1).

The literature of efficiency analysis has conventionally focused on fully parametric or nonparametric versions of model (1). Parametric models postulate a priori a specific functional form for *f* (e.g., Cobb-Douglas, translog, etc.) and subsequently estimate its unknown parameters. In contrast, axiomatic nonparametric models assume that *f* satisfies certain regularity axioms (e.g., monotonicity and concavity), but no particular functional form is assumed. At this point, we must emphasize that the term nonparametric does not necessarily imply that there are no restrictive assumptions. It is not true that the assumptions of a nonparametric model are necessarily less restrictive than those of a parametric model. For example, the fully nonparametric DEA estimator of model (1) is based on the assumption of no noise (i.e., $v_i = 0$ for all firms *i*). Assuming away noise does not require any specific parametric specification, but it is nevertheless a restrictive assumption. In fact, it is less restrictive to impose parametric structure and assume v_i are identically and independently distributed according to the normal distribution $N(0,\sigma_v^2)$. Note that this parametric specification contains the fully nonparametric "deterministic" case of no noise as its restricted special case, obtained by imposing the parameter restriction $\sigma_v^2 = 0$.

In addition to the pure parametric and nonparametric alternatives, the intermediate cases of semiparametric and semi-nonparametric models have become increasingly popular in recent years. However, the exact meaning of this terminology is often confused. Chen (2007) provides an intuitive and useful definition that we find worth quoting:

> "An econometric model is termed "*parametric*" if all of its parameters are in finite dimensional parameter spaces; a model is "*nonparametric*" if all of its parameters are in infinite-dimensional parameter spaces; a

> model is "*semiparametric*" if its parameters of interests are in finite-dimensional spaces but its nuisance parameters are in infinite-dimensional spaces; a model is "*semi-nonparametric*" if it contains both finite-dimensional and infinite-dimensional unknown parameters of interests". Chen (2007), p. 5552, footnote 1.

Note that according to the above definition both the semiparametric and semi-nonparametric model contain a nonparametric part and a parametric part. The distinction between the terms semiparametric and semi-nonparametric is subjective, dependent on whether we are interested in the empirical estimates of the nonparametric part or not. The same model can be either semiparametric, if our main interest is in the parameter estimates of the parametric part and the nonparametric part is of no particular interest, or semi-nonparametric, if we are interested in the results of the nonparametric part.

Model (1) can be interpreted as a neoclassical or frontier model depending on the interpretation of the disturbance term (cf., Kuosmanen and Fosgerau, 2009). The neoclassical model assumes that all firms are efficient and disturbances are random, uncorrelated noise terms. Frontier models typically assume that all or some part of the deviations from the frontier are attributed to systematic inefficiency.

Table 1 combines the criteria described above to identify six alternative estimation methods commonly used for estimating the variants of the unified model (1), together with some canonical references. On the parametric side, OLS refers to *ordinary least squares*, PP means *parametric programming*, COLS is *corrected ordinary least squares*, and SFA is *stochastic frontier analysis* (see, e.g., Kumbhakar and Lovell, 2000, for an introduction to the parametric approach to efficiency analysis). The focus of this chapter is on the axiomatic nonparametric and semi-nonparametric variants of model (1): CNLS refers to *convex nonparametric least squares* (Section 3), DEA is *data envelopment analysis* (Section 4.1), C^2NLS is *corrected convex non-parametric least squares* (Section 4.2), and StoNED is *stochastic nonparametric envelopment of data* (Section 5).

Table 1. Classification of methods

		Parametric	**Nonparametric**
Central tendency		*OLS* Cobb and Douglas (1928)	*CNLS* (Section 3) Hildreth (1954) Hanson and Pledger (1976)
Deterministic frontier	**Sign constraints**	*PP* Aigner and Chu (1968) Timmer (1971)	*DEA* (Section 4.1) Farrell (1957) Charnes et al. (1978)
	2-step estimation	*COLS* Winsten (1957) Greene (1980)	*C^2NLS* (Section 4.2) Kuosmanen and Johnson (2010)
Stochastic frontier		*SFA* Aigner et al. (1977) Meeusen and Vanden Broeck (1977)	*StoNED* (Section 5) Kuosmanen and Kortelainen (2012)

3. Convex nonparametric least squares

In this section we consider the special case of model (1) where the composite error term ε consists exclusively of noise v, and there is no inefficiency (i.e., we assume $u = 0$). This special case is relevant for modeling firms that operate in the competitive market environment, which meets (at least by approximation) the conditions of perfect competition considered in microeconomic theory. We will relax this no inefficiency assumption from Section 4 onwards, but the insights gained in this section will be critical for understanding the developments in the following sections.

In the case of a symmetric zero-mean error term that satisfies $E(\varepsilon_i) = 0$ for all i, the expected value of output conditional on inputs equals the value of the production function, that is,

$$E(y_i \,|\mathbf{x}_i) = E(f(\mathbf{x}_i)) + E(\varepsilon_i) = f(\mathbf{x}_i).$$

Therefore, in this setting the production function f can be estimated by nonparametric regression techniques. Note that the term "regression" refers to the conditional mean $E(y_i \,|\mathbf{x}_i)$.

Hildreth (1954) was the first to consider nonparametric regression subject to monotonicity and concavity constraints in the case of a single input variable x (see also Hanson and Pledger, 1976). Kuosmanen (2008) extended Hidreth's approach to the multivariate setting with a vector-valued $\mathbf{x}$, and coined the term *convex nonparametric least squares* (CNLS) for this method. CNLS builds upon the assumption that the true but unknown production function f belongs to the set of continuous, monotonic increasing and globally concave functions, F_2, imposing exactly the same production axioms as standard DEA.

The CNLS estimator of function f is obtained as the optimal solution to the infinite dimensional least squares problem

$$\begin{aligned} &\min_{f} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 \\ &\text{subject to} \\ &f \in F_2 \end{aligned} \qquad (2)$$

The functional form of f is not specified beforehand. Rather, the optimal solution will identify the best-fit function f from the family F_2. Note that set F_2 includes an infinite number of functions, which makes problem (2) impossible to solve through brute force trial and error. Further, problem (2) does not generally have a unique solution for any arbitrary input vector $\mathbf{x}$, but a unique solution exists for estimating f for the observed data points $(\mathbf{x}_i, y_i),\ i = 1,...,n$. Therefore, we will next discuss the estimation of f for the observed data points and extrapolation to unobserved points in sub-section 3.2.

3.1 CNLS estimator for the observed data points

A unique solution to problem (2) for the observed data points $(\mathbf{x}_i, y_i)$, $i = 1,...,n$, can be found by solving the following finite dimensional *quadratic programming* (QP) problem

$$\begin{aligned} &\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{CNLS})^2 \\ &\text{subject to} \\ &y_i = \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i + \varepsilon_i^{CNLS} \ \forall i \\ &\alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h' \mathbf{x}_i \ \forall h, i \\ &\boldsymbol{\beta}_i \ge 0 \ \forall i \end{aligned} \tag{3}$$

where α_i and $\boldsymbol{\beta}_i$ define the intercept and slope parameters of tangent hyperplanes that characterize the estimated piece-wise linear frontier (note that $\boldsymbol{\beta}_i'\mathbf{x}_i = \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + ... + \beta_{im}x_{im}$). Symbol ε_i^{CNLS} denotes the CNLS residual, which is an estimator of the true but unobserved $\varepsilon_i = v_i$. Note that in (3) the Greek letters are variables and the Latin letters are parameters (i.e., ($\mathbf{x}_i$, y_i) are observed data).

Kuosmanen (2008) introduced the QP formulation (3), and proved its equivalence with the infinite dimensional optimization problem (2). Specifically, if we denote the value of the objective function in the optimal solution to the infinite dimensional CNLS formulation (2) by SSE_{CNLS} (SSE = the sum of squares of errors), and that of the finite QP problem (3) by SSE_{QP}, then the equivalence can be stated as follows.

Theorem 1: $SSE_{CNLS} = SSE_{QP}$.

Proof. See Kuosmanen (2008), Theorem 2.1.

The equivalence result does not restrict to the objective functions, the optimal solution to problem (3) also provides us unique estimates of function *f* for the observed data points. Once the optimal solution is found, we will add "hats" on top of $\hat{\alpha}_i$, $\hat{\boldsymbol{\beta}}_i$, and $\hat{\varepsilon}_i^{CNLS}$, and refer to them as estimators.[7] In other words, α_i, $\boldsymbol{\beta}_i$, and ε_i^{CNLS} are variables of problem (3), whereas estimators $\hat{\alpha}_i$, $\hat{\boldsymbol{\beta}}_i$, and $\hat{\varepsilon}_i^{CNLS}$ provide the optimal solution to problem (3). Given $\hat{\alpha}_i$ and $\hat{\boldsymbol{\beta}}_i$ from (3), we define

$$\hat{f}^{CNLS}(\mathbf{x}_i) = \hat{\alpha}_i + \hat{\boldsymbol{\beta}}_i'\mathbf{x}_i = y_i - \hat{\varepsilon}_i^{CNLS}. \tag{4}$$

[7] In application, when estimators are calculated for a specific data set we will refer to these as estimated parameters.

This estimator of function *f* satisfies the following properties:

Theorem 2: *In the case of the neoclassical model with no inefficiency,* $\hat{f}^{CNLS}(\mathbf{x}_i)$ *is a unique, unbiased and consistent estimator of* $f(\mathbf{x}_i)$ *for the observed data points* $(\mathbf{x}_i, y_i),\ i = 1,...,n$.

Proof. Uniqueness is proved by Lim and Glynn (2012), Proposition 1. Unbiasedness follows from Seijo and Sen (2011), Lemma 2.4. Consistency is proved under slightly different assumptions in Seijo and Sen (2011), Theorems 3.1 and 3.2, and Lim and Glynn (2012), Theorems 1 and 2.

The constraints of the QP problem (3) have the following compelling interpretations.[8] The first constraint of the least squares formulation (3) is a linear regression equation. However, the CNLS regression does not assume linear *f*: note that coefficients α_i and $\boldsymbol{\beta}_i$ are specific to each observation *i*. Using the terminology of DEA, α_i and $\boldsymbol{\beta}_i$ are directly analogous to the multiplier coefficients of the dual formulation of DEA. The inequality constraints in (3) can be interpreted as a system of *Afriat inequalities* (compare with Afriat, 1967, 1972; and Varian, 1984). As Kuosmanen (2008) emphasizes, the Afriat inequalities are the key to modeling the concavity axiom in the general multiple regression setting.

Coefficients α_i and $\boldsymbol{\beta}_i$ should not be misinterpreted as parameters of the estimated function *f*, but rather, as parameters characterizing tangent hyperplanes to an unknown production function *f*. These coefficients characterize a convex piece-wise linear function, to be examined in more detail the next sub-section. At this point, we must emphasize that we did not assume or restrict the domain F_2 to only include piece-wise linear function. In fact, it turns out that the "optimal" functional form to solving the infinite dimensional least squares problem (2) is always a convex piece-wise linear function characterized by coefficients α_i and $\boldsymbol{\beta}_i$. However, this optimal solution is unique only for the observed data points.

3.2 Extrapolating to unobserved points

In many applications we are interested in estimating the frontier not only for the observed data points, but also for unobserved input vectors **x**. Although the CNLS estimator is unique for the observed data points, there is no unique way of extrapolating the CNLS estimator to unobserved points. In general, the optimal solution to the infinite dimensional least squares problem (2) is not unique, but there exists a set of functions $f^* \in F_2^*$ that solve the optimization problem (2). Formally, we denote the set of alternate optima to (2) as

[8] Note is formulation is written for ease of interpretation. Other formulations might be preferred to improve computational performance.

$$F_2^* = \left\{ f^* \,\middle|\, f^* = \arg\min_{f \in F_2} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 \right\}.$$

Kuosmanen (2008) characterizes the minimum and maximum bounds for the functions $f^* \in F_2^*$. It turns out that both bounds are piece-wise linear functions. However, only the minimum bound satisfies the postulated monotonicity and concavity properties. To resolve the non-uniqueness issue, Kuosmanen and Kortelainen (2012) appeal to the *minimum extrapolation principle* and propose to use the lower bound

$$\hat{f}_{\min}^{CNLS}(\mathbf{x}) = \min_{\alpha,\boldsymbol{\beta}} \left\{ \alpha + \boldsymbol{\beta}'\mathbf{x} \,\middle|\, \alpha + \boldsymbol{\beta}'\mathbf{x}_i \geq \hat{f}^{CNLS}(\mathbf{x}_i) \;\; \forall i = 1,\ldots,n \right\} \qquad (5)$$

Note that the lower bound $\hat{f}_{\min}^{CNLS}$ is simply the DEA estimator (single output, variable returns to scale) applied to the observed inputs $\mathbf{x}_i$ and the fitted outputs $\hat{f}^{CNLS}(\mathbf{x}_i)$ obtained from equation (4).[9] The lower bound function satisfies the postulated properties of monotonicity and concavity. We can make the following connection between the lower bound (5) and the infinite dimensional CNLS problem (2).

Theorem 3: *Function* $\hat{f}_{\min}^{CNLS}$ *stated in equation* (5) *is one of the optimal solutions to the infinite dimensional optimization problem* (2)*. It is the unique lower bound for the functions that solve problem* (2)*, formally*

$\hat{f}_{\min}^{CNLS}(\mathbf{x}) \leq f^*(\mathbf{x})$ for all $\mathbf{x} \in \Re_+^m$ and $f^* \in F_2^*$.

Proof. See Kuosmanen (2008) Theorem 4.1.

Note that while $\hat{f}^{CNLS}$ is unbiased and consistent for the observed points $\mathbf{x}_i$ (Theorem 3), the use of the piece-wise linear minimum function $\hat{f}_{\min}^{CNLS}$ will cause downward bias in finite samples as we apply the minimum extrapolation principle to extrapolate to unobserved points $\mathbf{x}$. Within the observed range of data, the downward bias will diminish as the sample size increases.

It is also worth noting that the optimal solution to the QP problem (3) does not necessarily produce unique coefficients $\hat{\alpha}_i$ and $\hat{\boldsymbol{\beta}}_i$. Although $\hat{f}_{\min}^{CNLS}$ is a unique lower bound, consistent with the minimum extrapolation principle, the coefficients $\hat{\alpha}_i$ and $\hat{\boldsymbol{\beta}}_i$ obtained as the optimal solution to (5) need not be unique either. It is well-known in the DEA literature that these multiplier coefficients are not unique in the vertices of the piece-wise linear function.

[9] In addition to the use of DEA to identify the lower bound function, there is a more fundamental connection between CNLS and DEA, to be explored in Section 4.

3.4 Computational issues

The CNLS problem (3) has linear constraints and a quadratic objective function, hence it can be solved by QCP solvers such as CPLEX or MOSEK.[10] Standard solvers work well in relatively small sample sizes (50 – 200 firms) available in the majority of published applications of efficiency analysis. However, since the number of Afriat inequalities in (3) grows at a quadratic rate as a function of the number of observations, the computational burden becomes a significant issue when the sample size increases beyond 300 firms. Note that adding a new firm to the sample increases the number of unknown parameters by m+2, and the number of Afriat inequality constraints increases by 2n. Introducing an additional input variable increases the number of unknown parameters by n, but there is no impact on the number of constraints. For these reasons, standard QP algorithms are inadequate for handling large samples with several hundreds or thousands of observations.

As a first step towards improving computational performance in small samples and to allow for larger problems to be solved, Lee et al. (2013) propose to follow the strategy of Dantzig *et al.* (1954, 1959) to iteratively identify and add violated constraints. The algorithm developed by Lee et al. first solves a relaxed CNLS problem containing an initial set of constraints, those that are likely to be binding, and then iteratively adds a subset of the violated concavity constraints until a solution that does not violate any constraint is found. In computational experiments, this algorithm allowed problems with up to 1,000 firms to be solved. Therefore, this algorithm has practical value especially in large sample applications and simulation-based methods such as bootstrapping or Monte Carlo studies. Another recent study by Hannah and Dunson (2013) implements CNLS in Matlab, reporting promising results. However, further algorithm development is needed to make the CNLS problem computable in very large sample sizes containing several thousands or millions of observations.

4. Deterministic frontiers

In this section we consider another special case of model (1) where the composite error term ε consists exclusively of inefficiency u, and there is no noise (i.e., $v = 0$). In the SFA literature, this special case is commonly referred to as the *deterministic model*. This does not imply, however, that probabilistic inferences are impossible.

Banker (1993) was the first to show that DEA can be understood as a maximum likelihood estimator of the deterministic model, with a statistical (probabilistic) foundation. However, the known statistical properties and inferences in the DEA literature restrict to the finite sample error that generally diminishes as the sample size increases. Or stated differently, the model

[10] Examples of computational codes for GAMS are available on the StoNED website: www.nomepre.net/stoned/.

specification and input and output data in the deterministic model are assumed to be exact and correct, so the only probabilistic component is the random sample of observations drawn from the production possibility set. This same deterministic model and its associated statistical foundation are used for inference in the bootstrapping methods (e.g., Simar and Wilson, 1998; 2000). Thus, statistical inference and confidence intervals estimated using bootstrapping methods only account for uncertainty in sampling and do not account for other sources of random variation or noise. Thus, bootstrap confidence intervals of DEA are not directly comparable to confidence intervals of other models that are genuinely stochastic in their nature (e.g., the SFA confidence intervals).

It is important to recognize that if the no noise assumption ($v = 0$) of the deterministic model does not hold, the statistical foundations of DEA collapse. The bootstrapping methods to adjust for the small sample are not a remedy against noise, rather adjusting for the sampling bias can make the DEA estimator worse if data are perturbed by noise. The stochastic case that includes both inefficiency and noise simultaneously will be considered in Section 5. The purpose of this section is to establish some useful connections between the 'neoclassical' CNLS and the 'deterministic' DEA to develop a unified framework and pave the way for a stochastic nonparametric StoNED estimator.

4.1 DEA as sign-constrained CNLS

In the single-output case, the variable returns to scale (VRS) DEA estimator of production function f can be stated as

$$\begin{aligned}\hat{f}^{DEA}(\mathbf{x}) &= \min_{\alpha,\boldsymbol{\beta}} \left\{ \alpha + \boldsymbol{\beta}'\mathbf{x} \middle| \alpha + \boldsymbol{\beta}'\mathbf{x}_i \geq y_i \ \ \forall i = 1,...,n \right\} \\ &= \max_{\boldsymbol{\lambda}} \left\{ \sum_{h=1}^{n} \lambda_h y_h \middle| \mathbf{x} \geq \sum_{h=1}^{n} \lambda_h \mathbf{x}_h ; \sum_{h=1}^{n} \lambda_h = 1 \right\}\end{aligned} \tag{6}$$

Note the difference between formulations (5) and (6): the former one uses the estimated output values $\hat{f}^{CNLS}(\mathbf{x}_i)$, whereas in the latter one uses the observed outputs y_i. Otherwise the formulations (5) and (6) are equivalent. The minimization formulation in (6) can be interpreted as the DEA multiplier formulation, whereas the maximization formulation of (6) is known as the DEA envelopment formulation. The duality theory of linear programming implies that the two formulations are equivalent.

Consider next a version of the CNLS estimator with an additional sign constraint on the residuals

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{CNLS-})^2$$

subject to

$$\begin{aligned} & y_i = \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i + \varepsilon_i^{CNLS-} \ \forall i \\ & \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h' \mathbf{x}_i \ \forall h,i \\ & \boldsymbol{\beta}_i \ge 0 \ \forall i \\ & \varepsilon_i^{CNLS-} \le 0 \ \forall i \end{aligned} \tag{7}$$

Comparing (3) and (7), we see that the only difference is the last constraint of (7), which is not present in the original CNLS formulation. Due to the sign constraint, Kuosmanen and Johnson (2010) interpret (7) as an axiomatic, nonparametric counterpart to the classic parametric programming approach of Aigner and Chu (1968).

We now establish the formal connection between CNLS and DEA as follows. Let $\hat{f}_{\min}^{CNLS-}(\mathbf{x})$ denote the piece-wise linear function obtained by applying equation (5) to the observed inputs $\mathbf{x}_i$ and the fitted values $\hat{y}_i$ of the sign-constrained formulation (7).

Theorem 4: *The sign-constrained CNLS estimator is equivalent to the DEA VRS estimator:*

$$\hat{f}_{\min}^{CNLS-}(\mathbf{x}) = \hat{f}^{DEA}(\mathbf{x})$$

Proof. Follows directly from Theorem 3.1 in Kuosmanen and Johnson (2010).

Although Theorem 4 was stated in the VRS case, the equivalence of DEA and sign-constrained CNLS does not restrict to the VRS case. Indeed parallel results are available for the other standard specifications of returns to scale by imposing additional constraints on the coefficients $\hat{\alpha}_i$ in formulations (3) or (7) as follows:

Constant returns to scale (CRS): impose $\hat{\alpha}_i = 0 \ \forall i$

Non-increasing returns to scale (NIRS): impose $\hat{\alpha}_i \ge 0 \ \forall i$

Non-decreasing returns to scale (NDRS): impose $\hat{\alpha}_i \le 0 \ \forall i$

Similarly, if the convexity assumption of DEA is relaxed the free disposable hull (FDH), Afriat (1972), estimator provides the minimum envelopment of data subject to free disposability. Keshvari and Kuosmanen (2013) show that the FDH formulation is a sign-constrained special case of isotonic nonparametric least squares (INLS), which in turn is the concavity relaxed version of CNLS.

From a practical point of view, the least squares interpretation of DEA opens up new avenues for applying tools from econometrics to DEA. For example, Kuosmanen and Johnson (2010) propose to measure the goodness-of-fit of DEA estimator by using the standard *coefficient of determination* from regression analysis, specifically

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}. \qquad (8)$$

Where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the average output in the sample. The R^2 statistic measures the proportion of output variation that is explained by the DEA frontier. While this variance decomposition can be applied to any regression model (including DEA), we note that DEA does not maximize the value of R^2 and hence negative R^2 values are possible for DEA estimators. This variance decomposition assumes a single output, however, one could compute and report separate R^2 statistics for each output.

4.2 Corrected CNLS

DEA builds on the minimum extrapolation principle to estimate the smallest function that envelops all data points. From the statistical point of view, insisting on the minimum extrapolation results in a systematic downward bias (i.e., the small sample error of DEA). For the deterministic model, Kuosmanen and Johnson (2010) show that a consistent and asymptotically unbiased estimator is obtained by applying a nonparametric variant of the classic COLS estimator. The proposed *corrected convex nonparametric least squares* (C^2NLS) estimator has always better discriminating power than DEA: the C^2NLS frontier envelops the DEA frontier everywhere, and the probability of finding multiple efficient units in randomly generated data approaches zero.

The C^2NLS method combines the nonparametric CNLS regression with the stepwise COLS approach first suggested by Winsten (1957), and more formally developed by Gabrielsen (1975) and Greene (1980). In this approach the most efficient firm in the sample is considered to be fully efficient, and the remaining inefficiency terms are normalized accordingly relative to the most efficient firm in the sample. A widely used panel data approach by Schmidt and Sickles (1984) applies a similar two-step approach (see Section 6.2 for details).

The essential steps of the C^2NLS routine can be described as follows:

Step 1: Apply the CNLS estimator (3) to estimate the conditional mean output $E(y_i \,|\mathbf{x}_i)$.

Step 2: Identify the most efficient unit in the sample (i.e., $\hat{u}_{benchmark}^{C2NLS} = \max_{h\in\{1,\ldots,n\}} \hat{\varepsilon}_h^{CNLS}$) as the benchmark. Adjust the CNLS residuals according to $\hat{u}_i^{C2NLS} = (\max_{h\in\{1,\ldots,n\}} \hat{\varepsilon}_h^{CNLS}) - \hat{\varepsilon}_i^{CNLS}$.

Step 3: Apply equation (5) to estimate the minimum function $\hat{f}_{\min}^{CNLS}(\mathbf{x})$. Adjust the minimum function by adding the residual of the benchmark firm to estimate the frontier using

$$\hat{f}^{C2NLS}(\mathbf{x}) = \hat{f}_{\min}^{CNLS}(\mathbf{x}) + \hat{u}_{benchmark}^{C2NLS}$$

Thus obtained $\hat{u}_i^{C2NLS}$ can be used as measures of inefficiency in the deterministic setting without noise. The most appealing properties of the C^2NLS estimator can be summarized as follows:

Theorem 5: *if* $\sigma_v = 0$*, then the* C^2*NLS estimator is statistically consistent:*

$$\operatorname{plim}_{n\to\infty} \hat{f}^{\mathrm{C2NLS}}(\mathbf{x}_i) = f(\mathbf{x}_i) \text{ for all } i = 1,\ldots,n.$$

Proof. Follows from Theorem 4.1 in Kuosmanen and Johnson (2010).

Theorem 6: *the* C^2*NLS frontier envelops the DEA frontier, that is,*

$$\hat{f}^{C2NLS}(\mathbf{x}) \geq \hat{f}^{DEA}(\mathbf{x}) \ \forall \mathbf{x} \in \Re_+^m.$$

Proof. Follows from Theorem 4.2 in Kuosmanen and Johnson (2010).

Note that the inefficiency estimates $\hat{u}_i^{C2NLS}$ are non-negative by construction, with the value of zero indicating full efficiency. The inefficiency measures can be converted to Farrell (1957) output efficiency scores ($\hat{\theta}_i^{C2NLS} \in [0,1]$) by using

$$\hat{\theta}_i^{C2NLS} = \frac{y_i}{\hat{f}^{C2NLS}(\mathbf{x}_i)} = \frac{y_i}{y_i + \hat{u}_i^{C2NLS}}. \qquad (9)$$

5. Stochastic Nonparametric Envelopment of Data (StoNED)

We are now equipped to consider the general stochastic nonparametric model that does not restrict to any particular functional form of *f* and includes both inefficiency *u* and stochastic noise *v*. Before proceeding to estimation, we must emphasize that the shift from the deterministic case to a stochastic model is

rather dramatic. For example, measuring the distance from an observed point to the frontier does not provide a measure of inefficiency if the observed point is perturbed by noise. While probabilistic inference in the deterministic case only investigates finite sample error, in the stochastic model the noise term is still relevant even if the sample size approaches infinity. Clearly, when all data points are subject to noise enveloping all observations would overestimate the true frontier production function. The CNLS regression that fits a monotonic increasing and concave curve through the middle of the cloud of data provides a natural starting point for the next generation of DEA that can deal with noise.[11] Following Kuosmanen (2006), we refer to this approach as *stochastic nonparametric envelopment of data* (StoNED).

Analogous to the parametric COLS and MOLS (*modified OLS*) estimators and the nonparametric C^2NLS, the StoNED estimator consists of multiple steps. The main steps can be described as follows (a detailed description of each step follows below):

Step 1: Apply the CNLS estimator (3) to estimate the conditional mean output $E(y_i \,|\mathbf{x}_i)$.

Step 2: Apply parametric methods (e.g., the method of moments or quasi-likelihood estimation) or nonparametric methods (e.g., kernel deconvolution) to the CNLS residuals ε_i^{CNLS} to estimate the expected value of inefficiency μ.

Step 3: Apply equation (5) to estimate the minimum function $\hat{g}_{\min}^{CNLS}(\mathbf{x})$. Adjust the minimum function by adding the expected inefficiency μ to estimate the frontier using

$$\hat{f}^{StoNED}(\mathbf{x}) = \hat{g}_{\min}^{CNLS}(\mathbf{x}) + \hat{\mu}$$

Step 4: Apply parametric methods (see e.g., Jondrow, Lovell, Materov and Schmidt, 1982, JLMS hereafter) or nonparametric deconvolution (e.g., kernel smoothing, Horrace and Parmeter, 2011) to estimate firm-specific inefficiency using the conditional mean $E(u_i \,|\varepsilon_i^{CNLS})$.

We will next describe each step in detail, noting that each step provides alternative modeling choices (depending on the assumptions one is willing to impose), and that it is not necessary to go through all of the steps. We discuss the information available at the end of each step and the possible motivations for proceeding to further steps.

[11] Banker and Maindiratta (1992) consider maximum likelihood estimation of the unified frontier model subject to monotonicy and concavity constraints. However, their maximum likelihood problem appears to be computationally prohibitive. We are not aware of any application of this method. Gstach (1998) presents another early attempt to incorporate noise in DEA. However, he needs to make a rather restrictive assumption of truncated noise (see Simar and Wilson, 2011, for sharp critique of this assumption).

5.1 Step 1: CNLS regression

The CNLS estimator was described in detail in Section 3 under the assumption of no inefficiency (u = 0). If the observed outputs are subject to asymmetric inefficiency, as the general frontier model (1) assumes, then the zero-mean assumption $E(\varepsilon_i)=0$ of regression analysis is violated. Indeed, $E(\varepsilon_i)=E(v_i-u_i)=-E(u_i)<0$ due to the asymmetric non-negative inefficiency term. Therefore, the CNLS estimator is no longer a consistent estimator of the frontier production function *f*.

Recall that CNLS regression estimates the conditional mean. Therefore, define the conditional mean function *g* as[12]

$$g(\mathbf{x}_i)=E(y_i\,|\,\mathbf{x}_i)=f(\mathbf{x}_i)-E(u_i)\,. \tag{10}$$

If the random inefficiency term u is independent of inputs $\mathbf{x}$, then the CNLS estimator $\hat{g}^{CNLS}(\mathbf{x}_i)$ is an unbiased and consistent estimator of function *g*. The CNLS estimator $\hat{g}^{CNLS}(\mathbf{x}_i)$ is obtained by solving the QP problem (3) and applying equation (4), as already discussed in Section 3, so we do not reproduce the CNLS formulations again here. Note that function *g* is simply the frontier production function *f* less the expected value of the inefficiency term u. If the inefficiency term u has a constant variance (i.e., inefficiency term u is homoscedastic), then the expected value of the inefficiency term u is a constant, denoted as μ. In other words, the CNLS provides a consistent estimator of the frontier *f* minus a constant. The constant μ can be estimated based on the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$, as discussed in more detail in Section 5.2. The case of heteroscedastic inefficiency where $E(u_i)$ is no longer a constant will be examined in Section 8.

Even if the data generating process (DGP) involves both inefficiency and noise, the CNLS estimator may be sufficient in some applications, without a need to proceed to the further stages. For example, if one is mainly interested in the relative efficiency rankings, then one could rank the evaluated units in descending order according to the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$. Further, if one is mainly interested in the marginal products of the input factors, the coefficients $\hat{\boldsymbol{\beta}}_i$ from (3), which are analogous to the multiplier coefficients (shadow prices) of DEA, then the CNLS regression provides consistent estimates (Seijo and Sen, 2011). The following steps described below do not influence the estimates of marginal products or the relative efficiency ranking of units. If one is interested in the frontier production function, average (in)efficiency in the sample, or cardinal firm-specific (in)efficiency estimates, then it is necessary to proceed further.

[12] Note that we use *g* to denote the conditional mean function when the composite error term contains inefficiency. This distinction was unnecessary in Section 3 because $g(x) = f(x)$ when there is no inefficiency present.

In the first step, one can impose some assumptions about returns to scale as described in Section 4.1. In addition, alternative modeling possibilities concern the multiplicative composite error and contextual variables are discussed as extensions in Section 6 and 7.

5.2 Step 2: Estimation of the expected inefficiency

Given the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$, it is possible to estimate the expected value of the inefficiency term $\mu = E(u_i)$. Note that if the variance of the inefficiency is constant across firms (the homoscedasticity assumption), then the expectation is taken unconditional and is constant across firms.

Alternative approaches for estimating μ are available. We will next briefly review the commonly used parametric approaches based on the method of moments (Aigner et al., 1977), quasi-likelihood estimation (Fan et al., 1996), and the nonparametric kernel deconvolution (Hall and Simar, 2002).

5.2.1 Method of moments

The method of moments requires some additional parametric distributional assumptions. The moment conditions are known at least for the commonly used half-normal and exponential inefficiency distributions, but not for all distributions considered in the SFA literature (e.g., the gamma distribution). In the following, we will discuss the commonly assumed case of half-normal inefficiency and normal noise. Stated formally, we assume

$$u_i \sim N^+(0, \sigma_u^2)$$

and

$$v_i \sim N(0, \sigma_v^2)$$

The CNLS residuals are known to sum to zero $\sum_{i=1}^{n} \hat{\varepsilon}_i^{CNLS} = 0$ (Seijo and Sen, 2011). Hence, we can calculate the second and the third central moment of the residual distribution as

$$\hat{M}_2 = \sum_{i=1}^{n} (\hat{\varepsilon}_i^{CNLS})^2 / (n-1) \qquad (11)$$

$$\hat{M}_3 = \sum_{i=1}^{n} (\hat{\varepsilon}_i^{CNLS})^3 / (n-1) . \qquad (12)$$

The second central moment $\hat{M}_2$ is simply the sample variance of the residuals and the third central moment $\hat{M}_3$ is a component of the skewness measure. The hats on top of these statistics indicate these statistics are estimators of the true but unknown values of the central moments. If the parametric assumptions of half-normal inefficiency and normal noise hold, then the second and the third central moments are equal to

$$M_2 = \left[\frac{\pi - 2}{\pi}\right]\sigma_u^2 + \sigma_v^2 \tag{13}$$

$$M_3 = \left(\sqrt{\frac{2}{\pi}}\right)\left[1 - \frac{4}{\pi}\right]\sigma_u^3 \tag{14}$$

Note that the third moment only depends on the standard deviation of the inefficiency distribution (σ_u). Thus, given the estimated $\hat{M}_3$ (which should be negative), we can estimate σ_u as

$$\hat{\sigma}_u = \sqrt[3]{\frac{\hat{M}_3}{\left(\sqrt{\frac{2}{\pi}}\right)\left[1 - \frac{4}{\pi}\right]}} \tag{15}$$

Subsequently, the standard deviation of the error term σ_v is estimated based on (12) as

$$\hat{\sigma}_v = \sqrt{\hat{M}_2 - \left[\frac{\pi - 2}{\pi}\right]\hat{\sigma}_u^2}\ . \tag{16}$$

There has been considerable discussion in the recent literature regarding the question of how to proceed if $\hat{M}_3$ is positive. Carree (2002), Alminidis et al. (2009), and Alminidis and Sickles (2012) consider alternative inefficiency distributions that allow for positive skewness. Simar and Wilson (2010) maintain the standard distributional assumptions, but suggest instead the use of bootstrapping method.

5.2.2 Quasi-likelihood estimation

Another way to estimate the standard deviations σ_u, σ_v is to apply the quasi-likelihood method suggested by Fan et al. (1996) (who refer to it as pseudo-likelihood). In this approach we apply the standard maximum likelihood (ML) method to estimate the parameters σ_u, σ_v, taking the shape of the CNLS curve as given (thus the term quasi-likelihood, in contrast to the full information ML which would also parameterize the coefficients of the frontier).

One of the main contributions of Fan et al. (1996) was to show that the quasi-likelihood function can be stated as a function of a single parameter (i.e., the signal-to-noise ratio $\lambda = \sigma_u / \sigma_v$)[13] as,

$$\ln L(\lambda) = -n \ln \hat{\sigma} + \sum_{i=1}^{n} \ln \Phi\left[\frac{-\hat{\varepsilon}_i \lambda}{\hat{\sigma}}\right] - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n} \hat{\varepsilon}_i^2\ , \tag{17}$$

[13] The signal-to-noise ratio λ should not be confused with the intensity weights λ_i used in the envelopment formulation of DEA.

where

$$\hat{\varepsilon}_i = \hat{\varepsilon}_i^{CNLS} - \left(\sqrt{2}\lambda\hat{\sigma}\right)\Big/\left[\pi\left(1+\lambda^2\right)\right]^{1/2}, \tag{18}$$

$$\hat{\sigma} = \left\{\frac{1}{n}\sum_{j=1}^{n}(\hat{\varepsilon}_i^{CNLS})^2 \Big/ \left[1 - \frac{2\lambda^2}{\pi(1+\lambda)}\right]\right\}^{1/2}. \tag{19}$$

Symbol Φ denotes the cumulative distribution function of the standard normal distribution N(0,1). We first use (18) and (19) to substitute out $\hat{\varepsilon}_i$ and $\hat{\sigma}$ from (17). We then maximize the quasi-likelihood function (17) by enumerating over λ values, using a simple grid search or more sophisticated search algorithms. When the quasi-likelihood estimate $\hat{\lambda}$ that maximizes (17) is found, we insert $\hat{\lambda}$ to equations (18) and (19) to obtain estimates of ε_i and σ. Subsequently, we can calculate estimates of $\hat{\sigma}_u = \hat{\sigma}\hat{\lambda}/(1+\hat{\lambda})$ and $\hat{\sigma}_v = \hat{\sigma}/(1+\hat{\lambda})$.

A simple practical trick to conduct quasi-likelihood estimation is to use ML algorithms available for SFA in standard software packages (e.g., Stata, Limdep, or R). By specifying the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ as the dependent variable (i.e., the output) and a constant term as an explanatory variable (input), we can trick the ML algorithm to perform the quasilikelihood estimation. This trick can also be used for estimating models involving contextual variables or heteroscedasticity (to be explored in Sections 7 and 8) by applying standard ML techniques as a second step.

5.2.3 Nonparametric kernel density estimation for the convoluted residual

While both method of moments and quasilikelihood techniques require parametric assumptions, a fully nonparametric alternative is available for estimating the signal-to-noise ratio λ, as proposed by Hall and Simar (2002). Their strategy is to search for a discontinuity in the residual density. The logic is that if an inefficiency term is left truncated, to represent efficient performance, there must be a discontinuity in distribution. When inefficiency is convoluted with noise, characterized by a continuous and smooth function, the discontinuity will still exist in the convoluted variable's density, the estimated residuals density. Thus, Hall and Simar suggest estimating the density of the residual using kernel methods and use these estimates to identify the largest change in the derivative on the right-side of the distribution (in the case of a production function and left-side in the case of the cost function). Then under the assumption of homoscedastic noise and inefficiency, the location of the largest change in the derivative can be used to estimate the mean inefficiency in the sample.

More formally, note that residuals $\hat{\varepsilon}_i^{CNLS}$ are consistent estimators of $\varepsilon_i^+ = \varepsilon_i + \mu$. Thus, we can apply the kernel density estimator for estimating the

density function of ε_i^+. Denote the kernel density estimator by f_{ε^+}. Hall and Simar (2002) show that the first derivative of the density function of the composite error term (f'_{ε}) is proportional to that of the inefficiency term (f'_u) in the neighborhood of μ. This is due to the assumption that f_u has a jump discontinuity at zero. Therefore, a robust nonparametric estimator of expected inefficiency μ is obtained as

$$\hat{\mu} = \arg\max_{z \in \Im} (\hat{f}'_{\varepsilon^+}(z)),$$

where $\Im$ is a closed interval in the right tail of f_{ε^+}.

5.3 Step 3: Estimating the frontier production function

In the presence of asymmetric inefficiency, the CNLS estimator estimates the conditional mean function $g(\mathbf{x}_i) = f(\mathbf{x}_i) - \mu$. Having estimated the expected inefficiency μ in Step 2, we can easily adjust the CNLS estimator to obtain an estimator of the frontier *f*. However, recall from Section 3 that the CNLS estimator of *g* is unique at the observed points $\mathbf{x}_i$ (i=1,…,n) but not in unobserved $\mathbf{x}$. Therefore, Kuosmanen and Kortelainen (2012) recommend applying the lower bound of *g* (analogous to equation (5)), defined as

$$\hat{g}^{CNLS}_{\min}(\mathbf{x}) = \min_{\alpha,\boldsymbol{\beta}} \left\{ \alpha + \boldsymbol{\beta}'\mathbf{x} \middle| \alpha + \boldsymbol{\beta}'\mathbf{x}_i \geq \hat{g}^{CNLS}(\mathbf{x}_i) \ \forall i = 1,...,n \right\}. \qquad (20)$$

We can subsequently add the expected inefficiency μ to estimate the frontier using

$$\hat{f}^{StoNED}(\mathbf{x}) = \hat{g}^{CNLS}_{\min}(\mathbf{x}) + \hat{\mu}.$$

This equation summarizes the relation between the StoNED frontier and the CNLS estimator as well as the relation between the frontier function *f* and the conditional mean function *g*. The heteroscedastic case where the shapes of the frontier *f* and the regression $E(y_i \,|\mathbf{x}_i)$ are different will be discussed in Section 8 below.

5.4 Step 4: Estimating firm-specific inefficiencies

Measuring the distance from an observation to frontier is not enough for estimating efficiency in the stochastic setting because all observations are subject to noise. Hence the measured distance to frontier consists of both inefficiency and noise (plus any error in our frontier estimate).

We must emphasize that even though there exist statistically unbiased and consistent methods for the estimation of the frontier *f*, there is no consistent method for estimating firm-specific efficiencies *u* in the cross-sectional setting subject to noise. In a cross-section, estimating firm-specific realizations of a

random variable u_i is impossible because we have only a single observation of each firm and all observations are perturbed by noise. This is not a fault of the methods (let alone their developers), it is just impossible to predict a realization of random variable based on a single observation that is subject to noise.

In the normal – half-normal case, Jondrow, Lovell, Materov and Schmidt (1982) (JLMS) develop a formula for the conditional distribution of inefficiency u_i given ε_i. The commonly used JLMS estimator for inefficiency is the conditional mean $E(u_i|\varepsilon_i)$. Given the parameter estimates $\hat{\sigma}_u$ and $\hat{\sigma}_v$, the conditional expected value of inefficiency can be calculated as[14]

$$E(u_i|\hat{\varepsilon}_i) = \frac{\hat{\sigma}_u\hat{\sigma}_v}{\sqrt{\hat{\sigma}_u^2+\hat{\sigma}_v^2}}\left[\frac{\phi\left(\frac{\hat{\varepsilon}_i\hat{\sigma}_u}{\hat{\sigma}_v\sqrt{\hat{\sigma}_u^2+\hat{\sigma}_v^2}}\right)}{1-\Phi\left(\frac{\hat{\varepsilon}_i\hat{\sigma}_u}{\hat{\sigma}_v\sqrt{\hat{\sigma}_u^2+\hat{\sigma}_v^2}}\right)}-\frac{\hat{\varepsilon}_i\hat{\sigma}_u}{\hat{\sigma}_v\sqrt{\hat{\sigma}_u^2+\hat{\sigma}_v^2}}\right], \quad (21)$$

where ϕ is the density function of the standard normal distribution $N(0,1)$, Φ is the corresponding cumulative distribution function, and

$$\hat{\varepsilon}_i = \hat{\varepsilon}_i^{CNLS} - \hat{\sigma}_u\sqrt{2/\pi}$$

is the estimator of the composite error term (compare with (18)). It is worth to note that there is nothing "stochastic" in the equation (21): the JLMS formula is a simply a deterministic transformation of the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ to a new metric that represents the conditional expected value of the inefficiency term. Indeed, the rank correlation of the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ and the JLMS inefficiency estimates is equal to one (see Ondrich and Ruggiero, 2001). For the purposes of relative efficiency rankings, the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ are sufficient.

Horrace and Parmeter (2011) show that the parametric assumption of the inefficiency distribution can be relaxed. Their approach still requires the parametric assumption of normally distributed noise. Rather than assuming a specific parametric distribution for the inefficiency term, the authors assume the density of *u* belongs to the ordinary smooth family of distributions, which includes exponential, gamma or Laplace (see also Fan, 1991). They apply Hall and Simar's (2002) method to estimate the jump discontinuity and thus the signal to noise ratio. Given the mean inefficiency level the authors are then able to construct the full density distribution of the inefficiency term using kernel smoothing and the residuals from a conditional mean estimation.

[14] Note that equation (21) is slightly different from the formula stated by Kuosmanen and Kortelainen (2012). Equation (21) is the corrected version stated by Keshvari and Kuosmanen (2013).

5.5 Statistical specification tests of the frontier model

As discussed above, the StoNED estimator consists of four steps. If all firms are efficient and deviations from the frontier are due to noise, the step 1 of estimating the conditional mean function is sufficient, and there is no reason to proceed further to step 2 of estimating the mean inefficiency to step 3 shifting the conditional mean function or step 4 estimating firm specific inefficiencies. To determine whether one should proceed from step 1 further to step 2, the efficiency analyst may want to test the data for evidence of inefficiency. If the results of a statistical specification test indicate that there is significant inefficiency present, this can be a convincing argument even for skeptics who believe that markets function efficiently.

The residual $\hat{\varepsilon}_i^{CNLS}$ consists of two components, a normally distributed noise term and a left-truncated inefficiency term. Schmidt and Lin (1984) propose a test of the skewness of the residuals as a method to investigate if inefficiency is present. By only looking at the skewness, the method is robust to the common alternative specifications of the inefficiency term in the stochastic frontier model. Thus, the null hypothesis is the residuals are normally distributed and a $\sqrt{b_1}$ test calculated as

$$\sqrt{b_1} = \frac{m_3}{(m_2)^{3/2}} \tag{22}$$

Where m_2 and m_3 are, the second and third moments of the residuals respectively. The distribution of the skewness test statistic, $\sqrt{b_1}$ can be constructed by a simple Monte Carlo simulation as described in D'Agostino and Pearson (1973). The authors also provide tables with critical values of the proposed test statistic for different sample sizes.

Kuosmanen and Fosgerau (2009) consider a fully nonparametric specification test that relaxes the normality assumption of the noise term. They show that the same test statistic $\sqrt{b_1}$ considered by Schmidt and Lin (1984) can be used for testing the null hypothesis of a symmetric v against the alternative hypothesis of skewness. They also recognize the $\sqrt{b_1}$ can wrongly reject the null hypothesis if the distribution is symmetric but has fat tails. Thus, they propose the additional b_2 test of the fourth moment

$$b_2 = \frac{m_4}{(m_2)^2} \tag{23}$$

Where m_2 and m_4 are the second and fourth moments of the residuals respectively. The null hypothesis is that the distribution is normally distributed. The alternative hypothesis is that there is non-normal kurtosis. The results of the $\sqrt{b_1}$ and b_2 tests can be given the following interpretation:

- If the null hypothesis of normality is rejected in the $\sqrt{b_1}$ test but maintained in the b_2 test, there is strong evidence in favor of a frontier model.

- If the null hypothesis of normality is maintained both in the $\sqrt{b_1}$ and b_2 tests, this supports the hypothesis of a competitive market with no inefficiency present.
- If the null hypothesis is rejected in the b_2 test, there may be data problems or model misspecification. There is no conclusive evidence in favor or against the frontier model.

It is worth noting that the power of the test depends on how specifically the null hypothesis and the alternative hypothesis are stated. For example, the $\sqrt{b_1}$ test of normality is more powerful than the fully nonparametric test of symmetry. If we are willing to impose some distributional assumptions for the inefficiency term, then more powerful specification tests are available. For example, Coelli (1995) proposed a variant of the Wald test to test the null hypothesis that there is no inefficiency, i.e. $\sigma_u^2 = 0$, against the alternative $\sigma_u^2 > 0$. While imposing distributional assumptions can increase the power of the test, it will also increase the risk of misspecification, which would make the statistical test inconsistent.

6. Extensions

6.1 Multiplicative composite error term

Most SFA studies use Cobb-Douglas or translog functional forms where inefficiency and noise affect production in a multiplicative fashion. In the present context, it is worth noting that the assumption of constant returns to scale (CRS) would also require multiplicative error structure, as will be discussed in more detail below. Further, a multiplicative error specification implies a specific model of heteroscedasticity in which the variance of the composite error term increases with firm size.

Multiplicative composite error structure is obtained by rephrasing model (1) as

$$y_i = f(\mathbf{x}_i) \cdot \exp(\varepsilon_i) = f(\mathbf{x}_i) \cdot \exp(v_i - u_i) \qquad (24)$$

Applying the log-transformation to equation (23), we obtain

$$\ln y_i = \ln f(\mathbf{x}_i) + \varepsilon_i \,. \qquad (25)$$

Note that the log-transformation cannot be applied directly to inputs **x** – it must be applied to the production function *f*.

In the multiplicative case, the CNLS formulation (3) can be rephrased as

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\phi,\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{CNLS})^2$$

subject to

$$\begin{aligned} &\ln y_i = \ln(\phi_i + 1) + \varepsilon_i^{CNLS} \quad \forall i \\ &\phi_i + 1 = \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i \quad \forall i \\ &\alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h' \mathbf{x}_i \quad \forall h,i \\ &\boldsymbol{\beta}_i \ge 0 \quad \forall i \end{aligned} \tag{26}$$

where $\phi_i + 1$ is the CNLS estimator of $E(y_i \,|\mathbf{x}_i)$. The value of one is added here to make sure that the computational algorithms do not try to take logarithm of zero. The first equality can be interpreted as the log transformed regression equation (using the natural logarithm function ln(.)). The second through fifth constraints are similar to (3) with the exception observed output in (3) is replaced with $\phi_i + 1$. The use of ϕ_i allows the estimation of a multiplicative relationship between output and input while assuring convexity of the production possibility set in original input-output space.[15]

Note that the log-transformation of a model variable renders the optimization formulation as a nonlinear programming (NLP) problem. These constraints are shown separately to illustrate the connection to previous formulations, but the first equality constraint can be moved to the objective function by solving and substituting for $\hat{\varepsilon}_i^{CNLS}$. Thus we have a convex solution space and a nonlinear objective function. This formulation can be solved by standard nonlinear programming algorithms and solvers. NLP solvers are available for example in such mathematical programming packages as GAMS, AIMMS, Matlab, and Lindo, among others.

In the multiplicative case, the CNLS estimator (25) can be applied, or as the first step of the C^{2}NLS or StoNED estimation routine. The standard method of moment, quasi-likelihood and kernel deconvolution techniques apply, as described in Section 5. However, note that in step 3 the frontier production function is obtained as $\hat{f}^{StoNED}(\mathbf{x}_i) = \hat{g}_{\min}^{CNLS}(\mathbf{x}) \cdot \exp(\hat{\mu})$, where $\hat{g}_{\min}^{CNLS}(\mathbf{x})$ is the minimum function computed using equation (5) and $\exp(\hat{\mu})$ is the estimated average efficiency. A convenient feature of the multiplicative model is that $\exp(u_i)$ can be interpreted as the Farrell output efficiency measure.

6.2 Panel data

In panel data the sample of firms is observed repeatedly over multiple time periods. Panel data applications are common in the SFA literature and a number of alternative SFA models involving time invariant and time varying inefficiency are available (see, e.g., Greene, 2008, Section 2.7). In contrast,

[15] If we apply the log transformation directly to input data, the resulting frontier would be a piece-wise log-linear frontier, which has been considered in the DEA literature by Charnes et al. (1982) and Banker and Maindiratta (1986). Unfortunately, the piece-wise log-linear frontier does not generally satisfy the concavity of *f*.

DEA studies ignore the time dimension of the panel data and either pool the panel together as a single cross section or treat each time period as an independent cross section.[16]

The regression interpretation of DEA examined in Section 4.1 allows us to combine DEA-style axiomatic frontier with the modern panel data methods from econometrics. Kuosmanen and Kortelainen (2012, Section 4.1) were the first consider a fixed effects approach to estimating a time invariant inefficiency model. Their fully nonparametric panel data StoNED estimator can be seen as a nonparametric counterpart to the classic SFA approach by Schmidt and Sickles (1984). In the following we consider the random effects approach, building upon Eskelinen and Kuosmanen (2013).

Consider a data set where each firm is observed over time periods $t = 1,...,T$ and define a time invariant frontier model

$$y_{it} = f(\mathbf{x}_{it}) - u_i + v_{it} \;\; \forall i = 1,...,n \;\; \forall t = 1,...,T \;, \tag{27}$$

where y_{it} is the observed output of firm *i* in time period *t*, $\mathbf{x}_{it}$ is a vector of inputs consumed by firm *i* in time period *t*, and *f* is a frontier production function that is time invariant and common to all firms. As before, u_i is a firm specific inefficiency term that does not change over time, and v_{it} is a random disturbance term of firm *i* in period *t*. Similar to the cross-sectional model, we assume that u_i and v_{it} are independent of inputs $\mathbf{x}_{it}$ and of each other.[17]

To estimate the model (27), we can adapt the standard CNLS estimator as

$$\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}} \sum_{t=1}^{T}\sum_{i=1}^{n} (\varepsilon_{it}^{CNLS})^2 \\
&\text{subject to} \\
&y_{it} = \alpha_{it} + \boldsymbol{\beta}_{it}'\mathbf{x}_{it} + \varepsilon_{it}^{CNLS} \;\; \forall i = 1,...,n \;\; \forall t = 1,...,T \\
&\alpha_{it} + \boldsymbol{\beta}_{it}'\mathbf{x}_{it} \le \alpha_{it} + \boldsymbol{\beta}_{it}'\mathbf{x}_{hs} \;\; \forall h,i = 1,...,n \;\; \forall s,t = 1,...,T \\
&\boldsymbol{\beta}_{it} \ge \mathbf{0} \;\; \forall i = 1,...,n \;\; \forall t = 1,...,T
\end{aligned} \tag{28}$$

where $\hat{\varepsilon}_{it}^{CNLS}$ is the CNLS residual of firm *i* in period *t*. Note the parameters α_{it} and $\boldsymbol{\beta}_{it}$ that define the tangent hyperplanes of the estimated production function are specific to each firm in each time period. Thus, a piece-wise linear frontier is estimated with as many as *nT* hyperplanes.

[16] One notable exception is Ruggiero (2004).

[17] The random effects approach to panel data requires that the time invariant inefficiency is uncorrelated with inputs. This is a strong assumption. Marschak and Andrews (1944) were among the first to note that rational firm manager will adjust the inputs to take into account the technical inefficiency, and hence the observed inputs are correlated with inefficiency. In that case, the random effects estimator is biased and inconsistent. The fixed effects estimator considered by Kuosmanen and Kortelainen (2012) does not depend on this assumption.

Given the optimal solution to (28), we compute the firm-specific effects as

$$\bar{\varepsilon}_i^{CNLS} = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_{it}^{CNLS} \tag{29}$$

Following Schmidt and Sickles (1984) we measure efficiency relative to the most efficient firm in the sample (analogous to the C^2NLS approach considered in Section 4.2) and define

$$\hat{u}_i^{StoNED} = (\max_{h\in\{1,\ldots,n\}} \bar{\varepsilon}_h^{CNLS}) - \bar{\varepsilon}_i^{CNLS}. \tag{30}$$

To estimate the conditional mean function, we can adapt equation (20) to panel data as

$$\hat{g}_{\min}^{CNLS}(\mathbf{x}) = \min_{\alpha,\boldsymbol{\beta}} \left\{ \alpha + \boldsymbol{\beta}'\mathbf{x} \middle| \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} \geq \hat{g}^{CNLS}(\mathbf{x}_{it}) \ \forall i = 1,\ldots,n; \forall t = 1,\ldots,T \right\}.$$

The StoNED frontier estimator is then obtained as

$$\hat{f}^{StoNED}(\mathbf{x}) = \hat{g}_{\min}^{CNLS}(\mathbf{x}) + (\max_{h\in\{1,\ldots,n\}} \bar{\varepsilon}_h^{CNLS}).$$

Both the frontier and inefficiency estimators can be shown to be statistically consistent under the assumptions stated above.

Note that the panel data StoNED estimator described above is fully nonparametric in the sense that no parametric functional form or distributional assumptions are required. Still, the model described in equation (27) relies on two strong assumptions: i) there is no technical progress, and ii) inefficiency is constant over time. It is possible to relax these assumptions, but this will require some additional assumptions (typically imposing some parametric structure). Note that random effects estimator considered above may still be useful even if inefficiency changes over time. In that case, the inefficiency estimator can be interpreted as the average efficiency during the time period under study. Eskelinen and Kuosmanen (2013) propose to examine the development trajectories of the normalized CNLS residuals $\hat{\varepsilon}_{it}^{CNLS} / (\max_{h\in\{1,\ldots,n\}} \bar{\varepsilon}_h^{CNLS})$ to gain a better understanding how the firm performance has developed during the study period. While the normalized CNLS residuals contain random noise, a growth trend (or decline) provides a clear indication that the performance of the firm has improved (or deteriorated) during the study period.

Based on the previous discussion, two insights are worth noting:

1) Panel data is not a panacea: while we recognize that panel data provides a richer set of information, we must also acknowledge that the intertemporal setting involves complex dynamics such as technological progress and changes in efficiency over time. The random effects approach to panel data considered above would be ideal for modeling experimental data where the researcher can control the input levels and keep the production technology the same across repeated experiments. However, most panel data applications of stochastic frontiers use observational data where both the production function and the level of efficiency will likely change over time.

2) Resorting to a fully nonparametric approach does not imply freedom from restrictive assumptions. In fact, avoidance of parametric assumptions often comes at the cost of very restrictive assumptions of no noise, no technical progress, or time invariant inefficiency. Indeed, insisting on a fully nonparametric approach can be more restrictive than resorting to some parametric assumptions that allow for explicit modeling of noise, technical progress, or time varying inefficiency.

6.3 Multiple outputs (DDF formulation)

The ability to model multiple inputs and multiple outputs has long been touted as an advantage of DEA over SFA: several DEA papers erroneously state that SFA cannot deal with multiple outputs. Lovell et al (1994) and Coelli and Perelman (1999; 2000) were the first to consider a stochastic distance function model that characterizes a general multiple inputs and multiple outputs technology using the radial input and output distance functions. The recent paper by Kuosmanen, Johnson and Parmeter (2013) (henceforth KJP) examines the assumptions of the data generation process that need to be satisfied for econometric identification of the distance function when the data are subject to random noise. Although the econometric estimation of distance functions is feasible, the well-established drawbacks of SFA still apply: a functional form needs to be specified for the distance function and parametric assumptions are typically made to decompose the residual into inefficiency and noise. Further, the commonly used parametric functional forms have the wrong curvature in output space, which is a serious problem for modeling joint production of multiple outputs.[18]

Up to this point, the CNLS/StoNED framework has been presented in the single output, multiple input setting. In this section we describe the CNLS estimator within the directional distance function (DDF) framework, Chambers

[18] The wrong curvature violates some of the most elementary properties of production technologies. For example, the Cobb-Douglas or translog specifications of the distance function will violates the basic properties of null jointness and unboundedness (see, e.g., Färe et al., 2005). Another problem concerns the economies of scope (e.g., Panzar and Willig, 1981). For example, the Cobb-Douglas distance function cannot capture the economies of scope at any parameter values. Since the economic rationale for joint production is rooted to economies of scope, it is contradictory to apply a technology that exhibits economies of specialization for modeling joint production.

et al. (1996, 1998). The CNLS formulation satisfies the axiomatic properties of the DDF by construction, models multiple inputs and multiple outputs, and accounts for stochastic noise explicitly, addressing the key limitations of both DEA and the parametric approaches. In the following we will briefly describe the stochastic data generating process (DGP) and the estimation of the DDF by CNLS. See KJP for a more detailed discussion.

The DDF indicates the distance from a given input-output vector to the boundary of the production possibility set T in some pre-assigned direction $(\mathbf{g}^x, \mathbf{g}^y) \in \Re_+^{m+s}$, formally,

$$\vec{D}_T(\mathbf{x}, \mathbf{y}, \mathbf{g}^x, \mathbf{g}^y) = \sup_\theta \left\{ \theta \middle| (\mathbf{x} - \theta \mathbf{g}^x, \mathbf{y} + \theta \mathbf{g}^y) \in T \right\}. \tag{31}$$

Denote the reference input-output vector of firm i in the direction $(\mathbf{g}^x, \mathbf{g}^y)$ by $(\mathbf{x}_i^*, \mathbf{y}_i^*)$. In this section we do not impose any particular behavioral hypothesis, but it may be illustrative to interpret $(\mathbf{x}_i^*, \mathbf{y}_i^*)$ as the optimal solution to firm i's profit maximization problem. Regardless of the firm manager's objective, we assume $(\mathbf{x}_i^*, \mathbf{y}_i^*)$ lies on the boundary of the production possibility set T and hence the values of the DDF satisfy

$$\vec{D}_T(\mathbf{x}_i^*, \mathbf{y}_i^*, \mathbf{g}^x, \mathbf{g}^y) = 0 \quad \forall i = 1, \ldots, n \tag{32}$$

The observed input-output vectors $(\mathbf{x}_i, \mathbf{y}_i)$, $i = 1, ..., n$, are perturbed in direction $(\mathbf{g}^x, \mathbf{g}^y) \in \Re_+^{m+s}$ by random inefficiency u_i and noise v_i, which form the composite error term $\varepsilon_i = u_i + v_i$ (note the positive sign of the inefficiency term u_i). Specifically, the observed data are perturbed versions of the optimal input-output vectors as follows

$$(\mathbf{x}_i, \mathbf{y}_i) = (\mathbf{x}_i^* + \varepsilon_i \mathbf{g}^x, \mathbf{y}_i^* - \varepsilon_i \mathbf{g}^y) \quad \forall i = 1, \ldots, n \tag{33}$$

We assume the inefficiency and noise terms satisfy the assumptions discussed in Section 2. Note that the elements of the direction vector $(\mathbf{g}^x, \mathbf{g}^y)$ represent the impacts of inefficiency and noise on specific input and output variables. If an element of $(\mathbf{g}^x, \mathbf{g}^y)$ is equal to zero, it means that the corresponding input or output variable is immune to both inefficiency and noise in the DGP. The larger the value of an element of $(\mathbf{g}^x, \mathbf{g}^y)$ in the DGP, the larger the impact of inefficiency and noise on the corresponding input or output variable is. Interestingly, Proposition 3 in KJP shows that in the DGP described above the value of the DDF equals the composite error term:

$$\overrightarrow{D}_T(\mathbf{x}_i,\mathbf{y}_i,\mathbf{g}^x,\mathbf{g}^y)=\varepsilon_i \quad \forall i\,.$$

This result provides implicitly a regression equation for estimating the DDF. We can resort to a similar stepwise procedure as described in Section 5.

The first step is to estimate the conditional mean distance defined as

$$d(\mathbf{x}_i,\mathbf{y}_i,\mathbf{g}^x,\mathbf{g}^y)=\overrightarrow{D}(\mathbf{x}_i,\mathbf{y}_i,\mathbf{g}^x,\mathbf{g}^y)-\mu \tag{34}$$

Let Δ denote the set of functions that satisfy the axioms of free disposability, convexity, and the translation property.[19] We can adapt the CNLS estimator to the DDF setting by postulating the following infinite dimensional least squares problem

$$\begin{aligned}&\min_{d}\sum_{i=1}^{n} d(\mathbf{x}_i,\mathbf{y}_i,\mathbf{g}^x,\mathbf{g}^y)^2\\ &\text{subject to}\\ &d\in\Delta\end{aligned} \tag{35}$$

Formulation (35) is a complex, infinite dimensional optimization problem that cannot be solved by brute-force numerical methods. The main challenge is to find a way to parameterize the infinitely large set of functions that satisfy the stated regularity conditions. Here again we apply insights from Kuosmanen (2008) and show an equivalent finite dimensional representation in terms of quadratic programming. Consider the following QP problem

$$\begin{aligned}&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\varepsilon}}\sum_{i=1}^{n}(\varepsilon_i^{CNLS})^2\\ &\text{subject to}\\ &\boldsymbol{\gamma}_i'\mathbf{y}_i=\alpha_i+\boldsymbol{\beta}_i'\mathbf{x}_i-\varepsilon_i^{CNLS}\ \ \forall i=1,\ldots,n\\ &\alpha_i+\boldsymbol{\beta}_i'\mathbf{x}_i-\boldsymbol{\gamma}_i'\mathbf{y}\le\alpha_h+\boldsymbol{\beta}_i'\mathbf{x}_i-\boldsymbol{\gamma}_h'\mathbf{y}_i\ \ \forall h,i=1,\ldots,n\\ &\boldsymbol{\gamma}_i'\mathbf{g}^y+\boldsymbol{\beta}_i'\mathbf{g}^x=1\ \forall i=1,\ldots,n\\ &\boldsymbol{\beta}_i\ge\mathbf{0}\ \forall i=1,\ldots,n\\ &\boldsymbol{\gamma}_i\ge\mathbf{0}\ \forall i=1,\ldots,n\end{aligned} \tag{36}$$

[19] The translation property, Chambers et al. (1998), states that if we move from the initial point $(\mathbf{x},\mathbf{y})$ in the direction $(\mathbf{g}^x,\mathbf{g}^y)$ by factor α, i.e., to the point $(\mathbf{x}+\alpha\mathbf{g}^x,\mathbf{y}-\alpha\mathbf{g}^y)$, then the distance to the frontier decreases by α. This property is crucial for the internal consistency of the DDF and can be seen as an additive analogue of the linear homogeneity property of the input distance function.

Note that the residual $\hat{\varepsilon}_i^{CNLS}$ here represents the estimated value of d_i (i.e., $\overrightarrow{D}(\mathbf{x}_i,\mathbf{y}_i,\mathbf{g}^x,\mathbf{g}^y)+u_i$). We also introduce new firm-specific coefficients $\boldsymbol{\gamma}_i$ that represent marginal effects of outputs to the DDF. The first constraint defines the distance to the frontier as a linear function of inputs and outputs. The linear approximation of the frontier is based on the tangent hyperplanes, analogous to the original CNLS formulation. The second set of constraints is the system of Afriat inequalities that impose global concavity. The third constraint is a normalization constraint that ensures the translation property. The last two constraints impose monotonicity in all inputs and outputs. It is straightforward to show that the CNLS estimator of function *d* satisfies the axioms of free disposability, convexity, and the translation property (see Theorem 3 in KJP).

After solving the CNLS problem, one can proceed to estimate the deterministic frontier by Corrected CNLS as described in Section 4.2 or the stochastic frontier by StoNED as described in Section 5.2. Note that the CNLS estimator described above does not estimate the DDF directly, but rather $\overrightarrow{D}\ (\mathbf{x}_i,\mathbf{y}_i,\mathbf{g}^x,\mathbf{g}^y)+E(u_i)$. If the inefficiency term is homoscedastic, then the techniques described in Section 5.2 apply for the estimation of $E(u_i)=\mu$. The case of heteroskedastic inefficiency term is discussed in Sections 8.2 and 8.3 below. Subsequently, the estimate of the DDF is obtained by shifting the CNLS estimate of function *d* in direction $(\mathbf{g}^x,\mathbf{g}^y)$ by the estimated expected inefficiency.

To connect the multi-output DDF to the single output case, it is worth noting in the single output case, specifying the direction vector as g^y=1 and $\mathbf{g}^x$=**0**, the CNLS problem (36) reduces to

$$
\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{CNLS})^2 \\
&\text{subject to} \\
&y_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i - \varepsilon_i^{CNLS} \quad \forall i = 1,\ldots,n \\
&\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \quad \forall h,i = 1,\ldots,n \\
&\boldsymbol{\beta}_i \ge \mathbf{0} \quad \forall i = 1,\ldots,n
\end{aligned}
\tag{37}
$$

This formulation is equivalent to the CNLS formulation (3) developed in Kuosmanen (2008), except for the sign of the residual $\hat{\varepsilon}_i^{CNLS}$ in the first constraint. Note that the DDF has positive values below the frontier and negative values above the frontier, which explains the negative sign.

6.4 Convex nonparametric quantile regression and percentile regression

While CNLS estimates the conditional mean $E(y_i\,|\,\mathbf{x}_i)$, quantile regression aims at estimating the conditional median or other quantiles of the response

variable (Koenker and Bassett, 1978; Koenker, 2005).[20] Denoting the pre-assigned quantile by parameter $q \in (0,1)$, we can modify the CNLS problem (3) to estimate convex nonparametric quantile regression (CNQR) (Wang et al., 2014) as follows:[21]

$$
\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}^+,\boldsymbol{\varepsilon}^-} q\sum_{i=1}^{n}\varepsilon_i^+ + (1-q)\sum_{i=1}^{n}\varepsilon_i^- \\
&\text{subject to} \\
&y_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i + \varepsilon_i^+ - \varepsilon_i^- \ \forall i \\
&\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \ \forall h,i \\
&\boldsymbol{\beta}_i \ge 0 \ \forall i \\
&\varepsilon_i^+ \ge 0 \ \forall i \\
&\varepsilon_i^- \ge 0 \ \forall i
\end{aligned}
\tag{38}
$$

The CNQR problem differs from CNLS in that the composite error term is now broken down to two non-negative components $\varepsilon_i^+, \varepsilon_i^- \ge 0$. The objective function minimizes the asymmetric absolute deviations from the frontier instead of symmetric quadratic deviations. The pre-assigned weight q defines the quantile to be estimated. For example, by setting q = 0.05, the piece-wise linear CNQR function will allow at most 5 percent of observations to lie above the fitted function and envelope at most 95 percent of the observed data points. As the sample size approaches to infinity, the q-order frontier will envelop exactly q percent of the observed data points (Wang et al., 2014, Theorem 1). Two important special cases are worth noting. First, if we set q = 0.5, then CNQR estimates the conditional median (whereas CNLS estimates the conditional mean). Secondly, as q approaches to zero, the negative deviations ε_i^- get a larger weight, and the CNQR approaches to the DEA frontier.

An appealing feature of the CNQR formulation is that its objective function and all constraints are linear functions of unknown parameters, and hence the CNQR problem can be solved by standard linear programming (LP) algorithms. However, a major drawback compared to CNLS is that the optimal solution to the CNQR problem is not necessarily unique, not even for the observed data points $(\mathbf{x}_i, y_i)$, $i = 1,...,n$. In econometrics, non-uniqueness of quantile regression is usually assumed away by assuming the regressors $\mathbf{x}$ are randomly drawn from a continuous distribution. In practice, however, input vectors $\mathbf{x}$ are not randomly drawn, and there may be two or more firms use exactly the same amounts of inputs (i.e., $\mathbf{x}_i = \mathbf{x}_j$ for firms i and j). In our experience, non-uniqueness of CNQR seems to be particularly a problem in

[20] In the DEA literature, the quantile frontiers are commonly referred to as robust order-*m* and order-*α* frontiers (e.g., Aragon et al. 2005; Daouia and Simar, 2007). However, while quantile frontiers are more robust to outliers than the conventional DEA frontiers, the quantile DEA approaches typically assume away noise.

[21] Similar quantile formulation was first considered by Banker et al. (1991), who refer to it as "stochastic DEA".

samples where inputs **x** are discrete variables. Wang et al. (2014) recognize non-uniqueness of the CNQR estimator, illustrating the problem with a numerical example.

One possible way to resolve the non-uniqueness problem is to apply the asymmetric least squares criterion suggested by Newey and Powell (1987), and reformulate the CNQR problem as

$$
\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}^+,\boldsymbol{\varepsilon}^-} q\sum_{i=1}^{n}(\varepsilon_i^+)^2+(1-q)\sum_{i=1}^{n}(\varepsilon_i^-)^2 \\
&\text{subject to} \\
&y_i=\alpha_i+\boldsymbol{\beta}_i'\mathbf{x}_i+\varepsilon_i^+-\varepsilon_i^- \ \forall i \\
&\alpha_i+\boldsymbol{\beta}_i'\mathbf{x}_i\le\alpha_h+\boldsymbol{\beta}_h'\mathbf{x}_i \ \forall h,i \\
&\boldsymbol{\beta}_i\ge 0 \ \forall i \\
&\varepsilon_i^+\ge 0 \ \forall i \\
&\varepsilon_i^-\ge 0 \ \forall i
\end{aligned}
\tag{39}
$$

To our knowledge, this asymmetric least squares formulation has not been considered before; we will henceforth refer to it as convex asymmetrically weighted least squares (CAWLS). The CAWLS problem differs from CNQR only in terms of the objective function, which now minimizes the asymmetric squared deviation instead of the absolute deviations. In the case of the linear regression, Newey and Powell (1987) show that the properties of the asymmetric least squares estimator are analogous to those of the quantile regression, but the asymmetric least squares can be more convenient for statistical inferences. In the present context, we hypothesize that the use of the quadratic loss function similar to CNLS ensures that the optimal solution to the CAWLS problem is always unique for the observed data points $(\mathbf{x}_i, y_i),\ i=1,...,n$. We leave confirming or rejecting this hypothesis as an open question for future research. Besides the question of uniqueness, the statistical properties of both CNQR and CAWLS would require further research.

CNQR and CAWLS formulations allow one to estimate the *q*-quantile or *q*-expectile frontiers directly, without a need to impose parametric distributional assumptions for the inefficiency and noise terms or resort to stepwise estimation along the lines described in Section 5. This is one of the attractive properties of CNQR and CAWLS. For the purposes of efficiency analysis, however, the use of quantiles or asymmetric weighted least squares is not a panacea. It is important to stress that the distance from the frontier, measured as $\hat{\varepsilon}_i^{CNQR}=\hat{\varepsilon}_i^+-\hat{\varepsilon}_i^-$ or $\hat{\varepsilon}_i^{CAWLS}=\hat{\varepsilon}_i^+-\hat{\varepsilon}_i^-$ (note: in both cases the residuals satisfy $\hat{\varepsilon}_i^+\hat{\varepsilon}_i^-=0\ \forall i$), should not be interpreted as a measure of inefficiency, as the distance to frontier also includes noise. To estimate conditional expected value of inefficiency along the lines of JLMS, we still need to resort to stepwise estimation. One possibility is to replace CNLS by CNQR or CAWLS as the first step of the StoNED procedure outlined in Section 5. Of course, residuals $\hat{\varepsilon}_i^{CNQR}$ or $\hat{\varepsilon}_i^{CAWLS}$ can be used as such for relative performance

rankings, but such performance rankings obviously depend on the chosen parameter value of q. Wang et al. (2014) examine the specification of q for frontier estimation, showing that the optimal value of q is a monotonically decreasing function of the signal to noise ratio $\lambda = \sigma_u / \sigma_v$. One may set the value of q based on subjective judgment, but in real world applications (consider, e.g., regulation of electricity distribution networks; see Kuosmanen, 2012; Kuosmanen, Saastamoinen and Sipiläinen, 2013), some objective criteria for specifying q would be important.

One appealing feature of the q-quantile and q-expectile frontiers is that they are robust to heteroscedasticity. Therefore, testing of and dealing with heteroscedasticity provide one promising application area for the CNQR and CAWLS techniques. If the composite error term is homoscedastic, then the quantile and expectile frontiers should have similar shapes at different values of q. Newey and Powell (1987) apply this idea for testing heteroscedasticity. We return to this issue in more detail in Section 8.

7. Contextual variables

A firm's ability to operate efficiently often depends on operational conditions and practices, such as the production environment and the firm specific characteristics for example technology selection or managerial practices. Banker and Natarajan (2008) refer to both variables that characterize operational conditions and practices as *contextual variables*. Currently two-stage DEA (2-DEA) is widely applied to investigate the importance of contextual variables as summarized by the citations included in Simar and Wilson (2007). However, its statistical foundation has been subject to sharp debate between Simar and Wilson (2007, 2011) and Banker and Natarajan (2008) (see also Hoff, 2007; McDonald, 2009). In this section we shed some new light on this debate following Johnson and Kuosmanen (2011, 2012).

It is important to note that Simar and Wilson (2007, 2011) do not consider stochastic noise in their DGP. In contrast, Banker and Natarajan (2008) introduce a noise term that has a doubly-truncated distribution, following the DEA+ approach by Gstach (1998). In this setting, Johnson and Kuosmanen (2012) show that the 2-DEA estimator of contextual variables is consistent under more general assumption that those stated by Banker and Natarajan (2008) and criticized by Simar and Wilson (2011). Further, Johnson and Kuosmanen (2012) employ the least squares formulation of DEA to develop a one-stage DEA method (1-DEA) for estimating the effects of the contextual variables. Relaxing the peculiar assumption of truncated noise,[22]

[22] We label this assumption as peculiar because it contradicts standard statistical assumptions, namely, the residual term is often model as normally distributed because a mixture of a large number of unknown distributions is approximately normal in finite samples and asymptotically normal. The large number of unknown distributions is a result of measurement errors, modeling simplifications, and other sources of noise. Thus, the motivation for truncated normal distribution used in Gstach (1998) and Banker and Natarajan (2008) is lacking and peculiar as also noted by Simar and Wilson (2011). Johnson and Kuosmanen (2012) argue this

Johnson and Kuosmanen (2011) develop *stochastic (semi-) nonparametric envelopment of z-variables data* (StoNEZD).

Taking the multiplicative model described in Section 6.1 as our starting point, we introduce the contextual variables, represented by *r*-dimensional vectors $\mathbf{z}_i$ that represent the measured values of operational conditions and practices, to obtain the following semi-nonparametric, partial log-linear equation

$$\ln y_i = \ln f(\mathbf{x}_i) + \boldsymbol{\delta}'\mathbf{z}_i + v_i - u_i . \qquad (40)$$

In this equation, parameter vector $\boldsymbol{\delta} = (\delta_1 \ldots \delta_r)'$ represents the marginal effects of contextual variables on output. All other variables maintain their previous definitions.

In the following sub-sections we will present two-stage DEA (2-DEA), one-stage DEA, and StoNEZD estimators. First, the 2-DEA estimator is described and the statistical properties of it are discussed. Given the assumptions necessary for the consistency of two-stage DEA method we then present the one-stage alternative. The joint estimation avoids the bias in the DEA frontier being transmitted to the parameter estimates of the coefficients on the contextual variables; however, the frontier estimated is still the minimum envelopment of the data and thus does not account for noise in the production model or input/output data. To account for stochastic noise, StoNEZD is introduced in 7.3.

7.1 Two-stage DEA

The literature on 2-DEA includes a number of variants. This sub-section follows the approach by Banker and Natarajan (2008). The two stages of their 2-DEA method are the following. In the first stage, the frontier production function f is estimated using the nonparametric DEA estimator formally stated as (5). The DEA output efficiency estimator of firm *i* is stated as $\hat{\theta}_i^{\mathrm{DEA}} = y_i / \hat{f}^{DEA}(\mathbf{x}_i)$ and computed as

$$(\theta_i^{\mathrm{DEA}})^{-1} = \max_{\theta\in\Re,\, \lambda\in\Re_+^n} \left\{ \theta \middle| \theta y_i \le \sum_{h=1}^n \lambda_h y_h ; \mathbf{x}_i \ge \sum_{h=1}^n \lambda_h \mathbf{x}_h ; \sum_{h=1}^n \lambda_h = 1 \right\} \qquad (41)$$

In the second stage, the following linear equation is estimated using OLS or ML

$$\ln \hat{\theta}_i^{\mathrm{DEA}} = \alpha + \boldsymbol{\delta}'\mathbf{z}_i + \varepsilon_i^{2-DEA},\ i = 1,...,n , \qquad (42)$$

---

truncation may come from an outlier detection procedure that would remove extreme observations from the analysis. However, in this case 1-DEA (introduced below) would still be preferred to 2-DEA because the bias introduced in two-stage estimation.

where the intercept α captures the expected inefficiency and the finite sample bias of the DEA estimator, and the composite disturbance term ε_i^{2-DEA} captures the noise term v_i and the deviations of u_i from the expected inefficiency μ. Note that the dependent variable has the "hat" because the DEA efficiency estimate is computed beforehand using (41), whereas the parameters on the right hand side of (42) are estimated using OLS or ML in a second stage.

Johnson and Kuosmanen (2012) state that the 2-DEA estimator is statistically consistent in the case of truncated noise as shown by Banker and Natarajan (2008), however, the assumptions required for consistency in Banker and Natarajan are unnecessarily restrictive.

Let $\mathbf{Z}$ denote a $n \times r$ matrix of contextual variables. Assume the noise terms are truncated as $|v_i| \le V^M$ and denote $\mathbf{v} = (v_1, ..., v_n)'$. Denote the domains of vectors $\mathbf{x}$ and $\mathbf{z}$ by D_x and D_z, respectively. Then the statistical consistency of the 2-DEA estimator can be established under the relaxed set of assumptions as follows.

Theorem 7: *If the following five assumptions are satisfied*

(i) *sequence* $\{(y_i, \mathbf{x}_i, \mathbf{z}_i)$, $i=1,...,n\}$ *is a random sample of independent observations,*

(ii) $\lim_{n\to\infty} \mathbf{Z}'\mathbf{Z}/n$ *is a positive definite matrix,*

(iii) *noise term* $\mathbf{v}$ *has a truncated distribution:* $|\mathbf{v}| \le V^M \mathbf{1}$, $f_v(V^M) > 0$,

(iv) *elements of domain* D_z *are bounded from above or below such that* $\boldsymbol{\delta}'\mathbf{z}$ *has a finite maximum* $\zeta = \max_{\mathbf{z}\in D_z} \boldsymbol{\delta}'\mathbf{z}$ *at a point* $\mathbf{z}^{\xi} \in \arg\max_{\mathbf{z}\in D_z} \boldsymbol{\delta}'\mathbf{z}$,

(v) *the joint density f is continuous and satisfies* $f(\mathbf{x}, \mathbf{z}^{\xi}, 0, V^M) > 0$ *for all* $\mathbf{x} \in D_x$,

then the 2-DEA estimators are statistically consistent in the following sense

$$\operatorname{plim}_{n\to\infty} \hat{f}^{\mathrm{DEA}}(\mathbf{x}_i) = f(\mathbf{x}_i) \cdot \exp(V^M + \zeta) \text{ for all } i = 1,...,n,$$

$$\operatorname{plim}_{n\to\infty} \hat{\boldsymbol{\delta}}^{\text{2-DEA}} = \boldsymbol{\delta}$$

Proof. See Johnson and Kuosmanen (2012), Theorem 1.

This theorem by Johnson and Kuosmanen (2012) generalizes the consistency result by Banker and Natarajan (2008) result by relaxing the following two assumptions:

1) inputs and contextual variables are statistically independent,
2) the effect of contextual variables is one-sided: $\mathbf{Z} \ge \mathbf{0}, \boldsymbol{\delta} \le \mathbf{0}$.

Note that the DEA frontier does not converge to the true frontier f, it converges to $f(\mathbf{x})\cdot\exp(V^M+\zeta)$ (i.e., the frontier augmented by the maximum noise V^M under the ideal conditions represented by $\mathbf{z}^\zeta$) thus estimation of the frontier requires observing firms that are operating efficiently and are operating in the best environment and happen to get a noise drawn close to the upper bound V^M.

Consistency is a relatively weak property. In practice a data set will be finite in size and probably not as large as we would like. However, Johnson and Kuosmanen (2012) are able to provide the explicit form of the bias in the 2-DEA estimator. Specifically it depends on the bias of the DEA frontier ($\hat{f}^{\text{DEA}}$) as follows:

$$\text{Bias}(\hat{\boldsymbol{\delta}}^{\text{2-DEA}}) = -(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\Big[\text{Bias}(\hat{f}^{\text{DEA}}(\mathbf{X}))\Big], \tag{43}$$

where

$$\text{Bias}(\hat{f}^{\text{DEA}}(\mathbf{X})) = \begin{pmatrix} E(\ln \hat{f}^{\text{DEA}}(\mathbf{x}_1)) - f(\mathbf{x}_1)\cdot\exp(V^M+\zeta) \\ \vdots \\ E(\ln \hat{f}^{\text{DEA}}(\mathbf{x}_n)) - \ln f(\mathbf{x}_n)\cdot\exp(V^M+\zeta) \end{pmatrix}.$$

Thus, the bias of the first-stage DEA estimator carries over to the second-stage OLS regression. Importantly, the bias of the second-stage OLS estimator is due to the correlation of $\mathbf{Z}$ and bias of the first-stage DEA estimator. In summary we would like to emphasize two critical points about 2-DEA.

> 1) correlation of inputs and contextual variables does not influence the statistical consistency of 2-DEA estimator as long as the columns of $\mathbf{X}$ and $\mathbf{Z}$ matrices are not linearly dependent.
> 2) the bias of the DEA frontier in the first-stage carries over to the second-stage OLS estimator through the correlation of the DEA frontier with the contextual variables.

We note that statistical independence of inputs and contextual variables does not necessarily guarantee that $\text{Bias}(\hat{f}^{\text{DEA}}(\mathbf{X}))$ is uncorrelated with $\mathbf{Z}$. Thus, 2-DEA does not suffer from some of the problems noted by Simar and Wilson (2011) and in fact requires significantly weaker assumptions than Banker and Natarajan (2008) suggest. However, the DEA frontier is always biased downward in a finite sample and thus this bias may be transferred to the estimation of the effect of the contextual variables. The following two sub-sections propose alternatives building on the regression interpretation of DEA which do not suffer from this bias.

7.2 One-stage DEA

The fundamental problem of the 2-DEA procedure is that the impact of the contextual variables $\mathbf{Z}$ is not taken into account in the first stage DEA. This problem has been recognized in the SFA literature, where the standard approach is to jointly estimate the frontier and the impacts of the contextual variables (e.g., Wang and Schmidt, 2002). In the similar vein, the least squares regression interpretation of DEA described in Section 4.1 allows us to estimate the DEA frontier and the coefficients $\boldsymbol{\delta}$ jointly. Specifically, we can introduce the contextual variables to the least squares formulation of DEA, stated as the QP problem (7), to obtain:

$$
\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta},\phi,\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{1-DEA})^2 \\
&\text{subject to} \\
&\ln y_i = \ln(\phi_i + 1) + \boldsymbol{\delta}'\mathbf{z}_i + \varepsilon_i^{1-DEA} \;\; \forall i \\
&\phi_i + 1 = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \;\; \forall i \\
&\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \;\; \forall h,i \\
&\boldsymbol{\beta}_i \ge 0 \;\; \forall i \\
&\varepsilon_i^{1-DEA} \le V^M \;\; \forall i
\end{aligned}
\tag{44}
$$

Notable differences compared to the problem (7) concern the use of the log-transformation to enforce the multiplicative formulation of the inefficiency term (compare with Section 6.1) and the truncation of the residual ε_i^{1-DEA} at point V^M . Note that by setting $V^M = 0$ restricts the noise term to zero, and the 1-DEA formulation reduces to the joint estimation of the effect of the contextual variables and the classic deterministic DEA frontier where all input/output data is observed exactly and residuals are non-positive.

Note further that the parameter vector $\boldsymbol{\delta}$ is common to all observations, and hence it can be harmlessly omitted from the Afriat inequalities that impose convexity. In fact, the contextual variables can be interpreted as inputs that have constant marginal products across all firms[23] (i.e., we can think of matrix $\mathbf{Z}$ as a subset of $\mathbf{X}$ for which $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j \;\; \forall i,j$).

The statistical properties of the 1-DEA estimator generally depend on the specification of the truncation point V^M . Performance of the 1-DEA estimator has been investigated via Monte Carlo simulations in Johnson and Kuosmanen (2012) where the authors find that 1-DEA performs well even when the truncation point is misspecified. However, the assumption of truncated noise (i.e., $|v_i| \le V^M$) is non-standard and debatable (see, e.g., Simar and Wilson, 2011). While the consistency of 2-DEA critically depends on this assumption, the CNLS estimator allows us to harmlessly relax it. The next sub-section

[23] This interpretation would vary slightly if the δ_i is negative. Then the contextual variable would be an output which would reduce the firm's ability to produce *y*.

discusses the StoNED estimator with z-variables that does not rely on the truncated noise assumption.

7.3 StoNED with z-variables (StoNEZD)

Relaxing the assumption of truncated noise, we can apply CNLS to jointly estimate the expected output conditional on inputs and the effects of the contextual variables. Johnson and Kuosmanen (2011) were the first to explore this approach, referring to it as StoNED with z-variables (StoNEZD). StoNEZD incorporates the contextual variables to the stepwise procedure sescribed in Section 5. In the following, we will focus on the CNLS estimator applied in the first step: steps 2 – 4 follow as described in Section 5, and are hence omitted here.

To incorporate the contextual variables in step 1 of the StoNED estimation routine, we can refine the multiplicative CNLS problem as follows:

$$
\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{\phi},\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{CNLS})^2 \\
&\text{subject to} \\
&\ln y_i = \ln(\phi_i + 1) + \boldsymbol{\delta}'\mathbf{z}_i + \varepsilon_i^{CNLS} \quad \forall i \\
&\phi_i + 1 = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \quad \forall i \\
&\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \quad \forall h,i \\
&\boldsymbol{\beta}_i \ge 0 \ \forall i
\end{aligned}
\tag{45}
$$

Note that problem (45) is identical to (44), except that the truncation constraint $\varepsilon_i \le V^M \ \forall i$ has been removed. Therefore, the least squares residuals are unrestricted, and hence problem (45) is a genuine conditional mean regression estimator.

Denote by $\hat{\boldsymbol{\delta}}^{StoNEZD}$ the coefficients of the contextual variables obtained as the optimal solution to (45). Johnson and Kuosmanen (2011) examine the statistical properties of this estimator in detail, showing its unbiasedness, consistency, and asymptotic efficiency.[24] Most importantly, the authors show that the conventional methods of statistical inference from linear regression analysis (e.g., t-tests, confidence intervals) can be applied for asymptotic inferences regarding coefficients $\boldsymbol{\delta}$. Their main result can be summarized as follows:

[24] Johnson and Kuosmanen (2012) report some Monte Carlo evidence of the finite sample performance of the StoNEZD estimator.

Theorem 8

If the following conditions are satisfied

i) *sequence* $\{(y_i, \mathbf{x}_i, \mathbf{z}_i)$, $i=1,...,n\}$ *is a random sample of independent observations,*

ii) $\lim_{n\to\infty} \mathbf{Z}'\mathbf{Z}/n$ *is a positive definite matrix,*

iii) *the inefficiency terms* $\mathbf{u}$ *and the noise terms* $\mathbf{v}$ *are identically and independently distributed (i.i.d.) random variables with* $Var(\mathbf{u}) = \sigma_u^2\mathbf{I}$ *and* $Var(\mathbf{v}) = \sigma_v^2\mathbf{I}$,

then the StoNEZD estimator for the coefficients of the contextual variables ($\hat{\boldsymbol{\delta}}^{StoNEZD}$ *) is statistically consistent and asymptotically normally distributed according to:*

$$\hat{\boldsymbol{\delta}}^{StoNEZD} \sim_a N(\boldsymbol{\delta}, (\sigma_v^2 + \sigma_u^2)(\mathbf{Z}'\mathbf{Z})^{-1}).$$

Proof. See Johnson and Kuosmanen (2011), Theorem 2.

This theorem extends the standard result of asymptotic normality of the OLS coefficients to the StoNEZD estimator of the contextual variables. In other words, even though model (40) includes a nonparametric function in addition to a linear regression function, the presence of the nonparametric function does not affect the limiting distribution of the parameter estimator in the linear part. In addition, Johnson and Kuosmanen (2011) show that the estimator $\hat{\boldsymbol{\delta}}^{StoNEZD}$ converges at the standard parametric rate, despite the presence of the nonparametric part in the regression equation. Therefore, we can apply the standard techniques from regression analysis such as *t*-tests and confidence intervals for asymptotic inferences.

A simple trick to compute standard errors for $\hat{\boldsymbol{\delta}}^{StoNEZD}$ is to run OLS regression where the contextual variables $\mathbf{Z}$ are regressors and the dependent variable is the difference between the natural log of observed output subtracting the natual log of the input aggregation plus 1, specifically $\ln y_i - \ln(\hat{\phi}_i + 1) = \hat{\boldsymbol{\delta}}'\mathbf{z}_i + \hat{\varepsilon}_i^{CNLS}$. This OLS regression will yield the same coefficients $\hat{\boldsymbol{\delta}}^{StoNEZD}$ that were obtained as the optimal solution to problem (45),[25] but also return the standard errors and other standard diagnostic statistics such as t-ratios, p-values, and confidence intervals.

[25] Note that this two-stage regression procedure is not subject to the problems of the 2-DEA procedure because we do control for the effects of the contextual variables in the first stage CNLS regression. It is just a computational trick to calculate the standard errors, but it can also serve as a simple diagnostic check that the solution to problem (32) is indeed optimal with respect to the contextual variables.

8. Heteroscedasticity

Up to this point we have assumed that the composite error term is homoscedastic, implying the variance parameters σ_u^2 and σ_v^2 are constant across all firms. This is a standard assumption both in regression analysis and in the parametric literature of frontier estimation (e.g., Aigner et al., 1977). However, this assumption is not always realistic in applications.

We can relax the assumption of constant σ_u^2 and σ_v^2, and allow these parameters to be firm specific (i.e., $\sigma_{u,i}^2$ and $\sigma_{v,i}^2$,), and potentially dependent on inputs **x** and contextual variables **z**. We stress that the least squares approach considered in this paper enables us to apply standard econometric techniques of testing and modeling heteroscedasticity considered in the SFA literature (see, e.g., Kumbkahar et al., 1991; Caudill and Ford, 1993; Caudill et al., 1995; Battese and Coelli, 1995; Hadri, 1999; and Kumbhakar and Lovell, 2000). The purpose of this section is to provide a brief review of how some of those techniques could be adapted for the purposes of CNLS and StoNED.

The first question to consider is how would heteroscedasticity affect the CNLS and StoNED estimators if we simply ignore it? Like standard OLS, the CNLS estimator remains unbiased and consistent despite heteroscedasticity. A weighted CNLS estimator (to be considered below) might be more efficient, provided that the heteroscedastic variance parameters can be estimated with a sufficient precision. However, heteroscedasticity is not a major problem for CNLS, and trying to improve its performance through explicit modeling and estimation of heteroscedasticity may not be worth the effort. Further research would be needed to investigate this issue.

The stepwise StoNED procedure is more sensitive to heteroscedasticity, as discussed by Kuosmanen and Kortelainen (2012). At this point, we need to distinguish between i) heteroscedastic inefficiency term and ii) heteroscedasticity noise term. Ignoring type ii) heteroscedasticity is less harmful in the StoNED estimation because the skewness of the CNLS residuals is still driven by the homoscedastic inefficiency term, the expected value of inefficiency is constant, and hence the shape of the regression function (i.e., the conditional mean $E(y_i|\mathbf{x}_i)$) is identical to that of the frontier production function *f*. Type i) heteroscedasticity will cause bigger problems, as Kuosmanen and Kortelainen (2012) recognize. If the inefficiency term is heteroscedastic, then the expected value of inefficiency is no longer constant, and the shapes of the regression function and the frontier production function will diverge. To take both types of heteroscedasticity explicitly into account, in Section 8.2 we will consider a doubly-heteroscedastic model where both inefficiency and noise terms are heteroscedastic. But before proceeding to the explicit modeling of heteroscedasticity, we describe a diagnostic test of the homoscedasticity assumption.

8.1 White test of heteroscedasticity applied to CNLS

Although the heteroscedastic inefficiency term would bias the StoNED estimator, it is important to emphasize that we do not need to take the homoscedasticity assumption by faith. Standard econometric tests of heteroscedasticity such as the White or the Breusch-Pagan tests are directly applicable to CNLS residuals. In this sub-section we briefly describe how the White (1980) test can be applied following Kuosmanen (2012).

The null hypothesis of the White test is that composite error term is homoscedastic, that is, H_0: $\sigma_{\varepsilon,i} = \sigma_{\varepsilon,j} \ \forall i, j$. The alternative hypothesis states there is heteroscedasticity, that is, H_1: $\sigma_{\varepsilon,i} \neq \sigma_{\varepsilon,j}$ for some *i,j*. Note that the alternative hypothesis does not assume any particular model of heteroscedasticity, which makes the White test compatible with the nonparametric approach. Postulating a more specific alternative hypothesis can increase the power of the test. However, the White test provides a useful starting point for more explicit modeling of heteroscedasticity.

The White test can be built upon the OLS regression of the following equation: [26]

$$(\hat{\varepsilon}_i^{CNLS})^2 = \alpha + \sum_{j=1}^{m} \beta_j x_{ij} + \frac{1}{2} \sum_{j=1}^{m} \sum_{h=1}^{j} \gamma_j x_{ij} x_{ih} + \varepsilon_i \,. \qquad (46)$$

In words, we explain the squared CNLS residual by a constant, all *m* input variables, and their squared values and cross-products using a flexible quadratic functional form as an approximation of the true but unknown heteroscedasticity effects. The test statistic is

$$W = nR^2 \,,$$

where R^2 is the coefficient of determination of the OLS regression of equation (46). Under the null hypothesis of homoscedasticity, the test statistic W follows the $\chi^2(J)$ distribution with *J* degrees of freedom, where $J = 1 + m + m(m+1)/2$ is the number of α, β, γ parameters on the right hand side of equation (46). If the value of test statistic W falls below the critical value of $\chi^2(J)$ at the given level of significance (note: the usual significance levels considered are 5% and 1%), then the null hypothesis of homoscedasticity is maintained. In that case, the test result provides some additional reassurance that the original model is well specified. On the other hand, if the value of test statistic W exceeds the critical value of $\chi^2(J)$ at the given level of significance, then the null

[26] In econometrics, heteroscedasticity is usually modeled as a function of explanatory variables (i.e., inputs **x**). In contrast, the SFA literature usually models heteroscedasticity as a function of **z**-variables that may contain some (or all) of the inputs **x**. For clarity, in this section we follow the econometric convention and focus on heteroscedasticity with respect to inputs **x** and discuss the additional **z**-variables below.

hypothesis is rejected, and hence explicit modeling of heteroscedasticity is needed.

The White test is usually presented in terms of the regressors of the original regression model (i.e., in terms of inputs **x** in the present context). Note that we are mainly concerned about possible heteroscedasticity with respect to inputs, which would cause bias in StoNED estimation. If we are interested in heteroscedasticity with respect to contextual variables **z**, we can also introduce the z-variables to the regression equation (46). We only need to adjust the degrees of freedom J to include the number of additional parameters for the z-variables, otherwise the test procedure is conducted as described above.

If significant heteroscedasticity is found, the White test does not indicate whether heteroscedasticity is in the inefficiency term or the noise term, or possibly both. To our knowledge, general diagnostic testing of whether heteroscedasticity is in the inefficiency or noise term has attracted little attention in the SFA literature. The doubly-heteroscedastic model (following Hadri, 1999; and Wang, 2002), to be examined in detail in the next sub-section, does allow us model heteroscedasticity in both inefficiency and noise terms, and also test for significance of the parameter estimates. However, such specification tests are conditional on the assumed model of heteroscedasticity, including the parametric distributional assumptions regarding inefficiency and noise. An appealing feature of the White test is it does not assume any specific model of heteroscedasticity and it does not depend on the distributional assumptions. Further, the parameter estimates of the auxiliary regression (46) and the associated diagnostic tools can provide some insights on which specific inputs (or contextual variables) are most likely causes of heteroscedasticity, and whether heteroscedasticity effect appears to be linear or non-linear, and whether the interaction terms (cross-products) are significant. These insights can be useful for specifying parametric models of heteroscedasticity, to be considered in the next sub-section.

Before proceeding, note that quantile estimation (see Section 6.4) could provide a promising nonparametric route for testing heteroscedasticity. If the composite error term is homoscedastic, then the q-quantiles should have approximately same shape for different values of parameter q. Provided that the number of input (and output) variables is sufficiently small, plotting the estimated q-quantiles at different values of q allow one to visually inspect whether homoscedasticity holds by a reasonable approximation. If homoscedasticity is violated, the q-quantile plots can help one to identify in which part of the frontier heteroscedasticity occurs, and which inputs are likely sources of heteroscedasticity. In the context of linear quantile regression, Koenker and Bassett (1982) propose formal tests of heteroscedasticity based on the comparison of the estimated q-quantiles at different values of q. Newey and Powell (1987) apply a similar idea for the q-expectiles, noting that the q-expectiles could also be used for testing symmetry of the composite error term (i.e., whether the asymmetric inefficiency term u is significant; compare with

Section 5.5). Adapting these tests to the nonparametric CNQR method for estimating q-quantiles and the CAWLS method for estimating q-expectiles introduced in Section 6.4 provides an interesting challenge for future research further discussed in section 9.

8.2 Doubly-heteroscedastic model

If the White test indicates significant heteroscedasticity, it is difficult to tell *a priori* whether heteroscedasticity is due to the inefficiency term, the noise term, or possibly both. Therefore, we will consider the general doubly-heteroscedastic model where both the inefficiency and noise term can be heteroscedastic. The doubly-heteroscedastic model was first considered by Hadri (1999). Our formulation below is mainly based on Wang (2002) and Kumbhakar and Sun (2013).

Consider the unified model described in Section 2. In this section we assume the inefficiency term has a truncated normal distribution and the noise term is normally distributed according to

$$u_i \sim N^+(\mu_i, \sigma_{u,i}^2)$$
$$v_i \sim N(0, \sigma_{v,i}^2)$$

The pre-truncation mean of the inefficiency term is assumed to be a linear function of inputs:

$$\mu_i = \alpha_0 + \boldsymbol{\beta}'\mathbf{x}_i .$$

The pre-truncation standard deviation of the inefficiency term and the standard deviation of the noise term are specified as

$$\sigma_{u,i} = \exp(\alpha_1 + \boldsymbol{\gamma}'\mathbf{x}_i)$$
$$\sigma_{v,i} = \exp(\alpha_2 + \boldsymbol{\rho}'\mathbf{x}_i)$$

Note that the exponent functions are commonly used in this context to guarantee that the standard deviations are positive at all input levels. While the specific parametric assumption may appear arbitrary, this model is one of the most flexible and general parametric specifications of heteroscedasticity. Note that the truncated normal distribution where both the pre-truncation mean and variance depend on the input level allows that the location (mean) and the shape (variance)of the inefficiency distribution can change as a function of inputs.

This formulation of heteroscedastic inefficiency term implies that the expected value of inefficiency can be stated as (see Wang, 2002; Kumbhakar and Sun, 2013)

$$E(u_i \mid u_i > 0) = \sigma_{u,i}\left[\Lambda_i + \frac{\phi(\Lambda_i)}{\Phi(\Lambda_i)}\right], \tag{47}$$

where

$$\Lambda_i = \frac{\mu_i}{\sigma_{u,i}}$$

and ϕ and Φ are the density function and the cumulative distribution function of the standard normal $N(0,1)$ distribution, respectively. The expected inefficiency is no longer a constant, but its dependence on inputs **x** has a well-defined functional form conditional on the parametric assumptions stated above. This allows us to both estimate the heteroscedasticity effects empirically, and take heteroscedasticity explicitly into account in the StoNED procedure.

8.3 Stepwise StoNED estimation under heteroscedasticity

To estimate the doubly-heteroskedastic model, we can adjust the stepwise StoNED routine presented in Section 5 as follows (a more detailed elaboration of each step follows below):

Step 1: Apply the CNLS estimator (3) to estimate the conditional mean output $\hat{g}^{CNLS}(\mathbf{x}_i) = E(y_i \mid \mathbf{x}_i)$.

Step 2: Apply quasi-likelihood estimation to the CNLS residuals ε_i^{CNLS} to estimate the parameters of μ_i, $\sigma_{u,i}$, and $\sigma_{v,i}$.

Step 3: Adjust the conditional mean function by adding the expected inefficiency $E(u_i \mid \mathbf{x}_i, \hat{\mu}_i, \hat{\sigma}_{u,i})$ to estimate the frontier for the observed data points using

$$\hat{f}^{StoNED}(\mathbf{x}_i) = \hat{g}^{CNLS}(\mathbf{x}_i) + E(u_i \mid \mathbf{x}_i, \hat{\mu}_i, \hat{\sigma}_{u,i}).$$

Then apply equation (5) to estimate the frontier $\hat{f}_{\min}^{StoNED}(\mathbf{x})$ for unobserved points.

Step 4: Apply JLMS method to estimate firm-specific inefficiency using the conditional mean $E(u_i \mid \hat{\varepsilon}_i^{CNLS})$.

In step 1, we estimate the conditional mean function $g(\mathbf{x})$. The CNLS estimator remains unbiased and consistent estimator of the conditional mean g, despite heteroscedastic composite error term (similar to OLS). However, note that in the case of the doubly-heteroscedastic model

$$g(\mathbf{x}_i) = E(y_i \mid \mathbf{x}_i) = f(\mathbf{x}_i) - E(u_i \mid \mathbf{x}_i).$$

Note that the shape of function g can differ from that of frontier f because $E(u_i|\mathbf{x}_i)$ is a function of inputs $\mathbf{x}$. We will take this into account in step 3 where we shift function g upward, not by a constant μ, but rather, by the estimated $E(u_i|\mathbf{x}_i)$.[27] It is also worth noting that function g is not necessarily monotonic increasing and concave even if the production function f satisfies these axioms because $-E(u_i|\mathbf{x}_i)$ can be a non-monotonic and non-concave function of inputs (note: there does exist parameter values for which $-E(u_i|\mathbf{x}_i)$ is indeed monotonic and concave in the domain of non-negative $\mathbf{x}$). To apply CNLS in step 1, we need to assume that the curvature of the production function f dominates and that function g is monotonic increasing and concave (at least by approximation). Even if one assumes that f exhibits CRS, it is recommended to apply the VRS specification in step 1 to allow for the nonlinear effects of $E(u_i|\mathbf{x}_i)$, and impose CRS later in step 3.

Having estimated the parameters of the inefficiency and noise terms, it is possible to test if monotonicity and concavity assumptions of g hold. If g does not satisfy monotonicity and concavity, we can substitute CNLS by techniques depending on which axiom does not hold. Specifically, if the concavity assumption is violated, it is possible to apply isotonic nonparametric least squares (INLS) suggested by Keshvari and Kuosmanen (2013). Another possibility is to estimate order-q quantile frontier using either CNQR or CAWLS techniques introduced in Section 6.4. Specifying the correct value for q will ensure that the quantile frontier inherits the monotonicity and concavity properties of frontier f even if the heteroscedastic inefficiency term is a non-monotonic or non-convex function of inputs. Indeed, we do not insist on estimating the conditional mean in step 1, the conditional quantile is equally suitable.

In step 2 it is natural to resort to the pseudolikelihood method since we utilize a rather heavily parametrized model of heteroscedasticity. As already noted in Section 5, a simple practical trick to conduct quasi-likelihood estimation is to use the standard ML algorithms available for SFA in standard software packages (e.g., Stata, Limdep, or R). In this case we specify the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ as the dependent variable (i.e., the output) and a constant term as an explanatory variable (input), and the ML algorithm performs the quasilikelihood estimation. For example, the frontier modeling tools of Stata allows one to include "explanatory variables for technical inefficiency variance function (uhet)" and "explanatory variables for idiosyncratic error variance function (vhet)" if the distribution of inefficiency term is specified as half-

[27] In the context of SFA, Kumbhakar and Lovell (2000) state strongly that the stepwise MOLS procedure cannot be used in the case of heteroscedastic inefficiency. They correctly note that the OLS estimator used in the first step yields biased estimates of not only the intercept but also the slope coefficients of the frontier. However, Kumbhakar and Lovell seem to overlook the possibility of eliminating the bias by shifting function g upward by a conditional expectation of inefficiency that depends on inputs $\mathbf{x}$.

normal or exponential. It is also possible to include covariates to the truncated normal specification of the inefficiency term, but in this specification the noise term is assumed to be homoscedastic. Hung-Jen Wang has developed a Stata package for the model described in Wang (2002), which can be used for estimating the model estimating the heteroscedasticity model described above.[28]

Having estimated the underlying parameters of $\mu_i, \sigma_{u,i}, \sigma_{v,i}$, it is recommended to apply standard specification tests available for ML (i.e., likelihood-ratio, Lagrange multiplier, or Wald test) to test restrictions $\boldsymbol{\beta} = \mathbf{0}$, $\boldsymbol{\gamma} = \mathbf{0}$, and $\boldsymbol{\rho} = \mathbf{0}$. For example, if the null hypothesis of $\boldsymbol{\rho} = \mathbf{0}$ is not rejected, then the assumption of homoscedastic noise term can be maintained. Similarly, if $\alpha_0 = 0$, $\boldsymbol{\beta} = \mathbf{0}$, and $\boldsymbol{\gamma} = \mathbf{0}$, then the model of heteroscedastic truncated normal inefficiency term reduces to a homoscedastic half-normal inefficiency term. If the specification tests provide evidence that some of the heteroscedasticity effects are not significant, we would recommend excluding those effects from the heteroscedasticity model and estimating step 2 again.

One additional issue is in the context of linear regressionthat efficiency of the least squares estimator can be improved by applying weighted least squares or generalized least squares. Having estimated the firm specific $\sigma_{u,i}, \sigma_{v,i}$, it is possible to return back to step 1 and apply a weighted version of the CNLS estimator. Defining $\hat{\sigma}^2_{\varepsilon,i} = \hat{\sigma}^2_{u,i} + \hat{\sigma}^2_{v,i}$, we can modify the objective function of the CNLS problem as

$$\min \sum_{i=1}^{n} \frac{(\varepsilon_i^{CNLS})^2}{\hat{\sigma}^2_{\varepsilon,i}}$$

maintaining the original constraints of (3). Interpreting the given $1/\hat{\sigma}^2_{\varepsilon,i}$ as firm-specific weights, this weighted least squares formulation of CNLS is directly analogous to the generalized least squares (GLS) estimator of the linear regression model.[29] However, as yet there is no evidence that the use of weighted least squares can improve efficiency of the CNLS estimator. Intuitively, the direct analogue with GLS would suggest that weighted least squares can be more efficient than the unweighted CNLS under heteroscedasticity. On the other hand, recall that CNLS approximates the underlying function *g* by a piece-wise linear curve. Since the hyperplane segments of the unweighted CNLS formulation provide local approximation, assigning larger or smaller weights to certain regions of the frontier may not have much effect on the piece-wise linear approximation. In our limited experience, introducing the weights $1/\hat{\sigma}^2_{\varepsilon,i}$ does not necessarily have any

[28] The Stata package is available from Wang's homepage: http://homepage.ntu.edu.tw/~wangh/.

[29] Note that in the CNLS context we prefer to introduce weights to the objective function instead of applying variable transformations (as in GLS) because the monotonicity and concavity constraints must hold for the original input variables **x**.

notable impact on the results. Further, we need to be able to estimate $\sigma^2_{\varepsilon,i}$ with a sufficient precision. Overall, we are somewhat skeptical whether the possible benefit in terms of improved efficiency of the CNLS estimator can outweigh the cost of additional effort of conducting the weighted least squares estimation. This forms an interesting open question for future research.

In step 3 we adjust the conditional mean function *g* estimated in step 1 (or alternatively, the conditional *q*-quantile) for the estimated expected inefficiency to estimate the frontier *f*. Note that the conditional mean $E(u_i|\mathbf{x}_i)$ is no longer a constant, but a function that depends on inputs **x**. Using equation (47), we can write the estimated expected inefficiency as the function of inputs and parameter estimates as

$$E(u_i|\mathbf{x}_i,\hat{\mu}_i,\hat{\sigma}_{u,i}) = \hat{\mu}_i + \hat{\sigma}_{u,i}\frac{\phi(\hat{\Lambda}_i)}{\Phi(\hat{\Lambda}_i)}$$
$$= (\hat{\alpha}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}_i) + \exp(\hat{\alpha}_1 + \hat{\boldsymbol{\gamma}}'\mathbf{x}_i)\left[\phi\left(\frac{\hat{\alpha}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}_i}{\exp(\hat{\alpha}_1 + \hat{\boldsymbol{\gamma}}'\mathbf{x}_i)}\right)\Big/\Phi\left(\frac{\hat{\alpha}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}_i}{\exp(\hat{\alpha}_1 + \hat{\boldsymbol{\gamma}}'\mathbf{x}_i)}\right)\right]$$

This expression reveals that in the doubly-heteroscedastic model the expected value of inefficiency has a linear part originating from the mean $\mu_i = \alpha_0 + \boldsymbol{\beta}'\mathbf{x}_i$, and a nonlinear part driven by $\sigma_{u,i} = \exp(\alpha_1 + \boldsymbol{\gamma}'\mathbf{x}_i)$. Having estimated the parameters of the inefficiency term, it is useful to evaluate whether $-E(\hat{u}_i|\mathbf{x}_i)$ is monotonically increasing and concave within the observed range of inputs (e.g., plot the values of $-E(\hat{u}|\mathbf{x})$ at different levels of **x** to visually inspect possible violations of monotonicity and concavity). To ensure that the estimated frontier function satisfies the postulated axioms despite minor violations of monotonicity and concavity (which may be just artifacts of the arbitrary parametric specification of the heteroscedasticity model), we apply the minimum extrapolation principle and utilize the DEA method stated in equation (5) to obtain the convex monotonic hull of the fitted values $\hat{f}^{StoNED}(\mathbf{x}_i)$ of observations i=1,…,n, which yields the frontier estimator $\hat{f}^{StoNED}_{\min}(\mathbf{x})$.

In step 4, we can compute firm specific inefficiency estimates using the JLMS conditional mean $E(u_i|\hat{\varepsilon}_i^{CNLS})$ using the firm specific parameter estimates $\hat{\mu}_i, \hat{\sigma}_{u,i}, \hat{\sigma}_{v,i}$. Note that the expected inefficiency $E(u_i|\mathbf{x}_i,\hat{\mu}_i,\hat{\sigma}_{u,i})$ applied for shifting the conditional mean function *g* to estimate frontier *f* does not depend on the heteroscedasticity of the noise term. However, the JLMS efficiency does also depend on the heteroscedasticity of the noise term $\hat{\sigma}_{v,i}$. Kumbhakar and Sun (2013) discuss this issue in more detail, showing that the

marginal effect of inputs on the conditional JLMS efficiency also depend on the heteroscedasticity of the noise term.

9. Directions for future research

This chapter has provided an updated and elaborated presentation of the CNLS and StoNED methods. Bridging the gap between the established DEA and SFA paradigms, these methods represent a major paradigm shift towards a unified and integrated methodology of frontier estimation and efficiency analysis that has a considerably broader scope than the conventional DEA and SFA tools. This chapter did not only review previously published method developments and their extensions, but also presented some new innovations, including the first extension of the StoNED method to the general case of multiple inputs and multiple outputs, and the first detailed examination of how heteroscedastic inefficiency and noise terms can be modeled within the CNLS and StoNED estimation frameworks.

We see CNLS and StoNED not only as the state of the art in axiomatic nonparametric frontier estimation and efficiency analysis under stochastic noise, but also a promising way forward. Kuosmanen and Kortelainen (2012) stated explicitly 12 promising avenues of future research on the StoNED methodology. In the following we will provide an updated version of a 12 point research program, indicating the work that has already been done as well as work that remains to be done.

”1. *Adapting the known econometric and statistical methods for dealing with heteroskedasticity, endogeneity, sample selection, and other potential sources of bias, to the context of CNLS and StoNED estimators*.”
In this chapter we presented the first detailed examination about the modeling of heteroscedasticity in the inefficiency and noise terms. Kuosmanen, Johnson and Parmeter (2013) examine the endogeneity problem from a novel perspective employing directional distance functions. Obviously, a lot of further work is needed in this area. Alternative models of heteroscedasticity as well as estimation techniques deserve careful attention. The convex nonparametric quantile regression and the convex asymmetrically weighted least squares methods discussed in Section 6.4 and the generalized least squares estimator discussed in Section 8.3 provide potential methods for modeling and testing heteroskedasticity. The use of instrumental variables in CNLS for modeling measurement errors, sample selection, and other types of endogeneity bias should be investigated.

“2. *Extending the proposed approach to a multiple output setting.*”
In this chapter we also presented the first extension of the StoNED method to the general case of multiple inputs and multiple outputs using the directional distance function (see also Kuosmanen, Johnson and Parmeter, 2013). Further work is also needed in this area. Alternative representations of the joint

production technology, including the radial input and output distance functions, should be investigated. The main challenge in modeling joint production is not the formulation of the mathematical programming problem for the CNLS estimator (the usual DEA problem) or deconvoluting the composite error term (the usual SFA problem). The main challenge is the probabilistic modeling of the data generating process in the case of joint production, involving multiple endogenous inputs and outputs. Kuosmanen, Johnson and Parmeter (2013) provides a useful starting point in this respect.

"3. *Extending the proposed approach to account for relaxed concavity assumptions (e.g., quasiconcavity)*."
Keshvari and Kuosmanen (2013) presented the first extension in this direction, applying isotonic regression that relaxes the concavity assumption of CNLS. This approach estimates a step function analogous to free disposable hull (FDH) in the middle of the data cloud. The insights of Keshvari and Kuosmanen could be useful for examining the intermediate cases between the non-convex step function and the fully convex CNLS, allowing one to postulate quasiconcavity or quasiconvexity in terms of some variables (e.g., inputs, or input prices in the estimation of the cost function). Many opportunities for future research exist in this direction.

"4. *Developing more efficient computational algorithms or heuristics for solving the CNLS problem*."
Lee et al. (2013) is the first contribution in this direction. The algorithm developed in that paper first solves a relaxed CNLS problem containing an initial set of constraints, those that are likely to be binding, and then iteratively adds a subset of the violated concavity constraints until a solution that does not violate any constraint is found. We believe the computational efficiency can be improved considerably by clever algorithms and heuristics (see, e.g., Hannah and Dunson, 2013). This is an important avenue for future research in the era of "big data".

"5. *Examining the statistical properties of the CNLS estimator, especially in the multivariate case*."
Seijo and Sen (2011) and Lim and Glynn (2012) were the first to address this challenge, proving statistical consistency of the CNLS estimator in the general multivariate case under slightly different assumptions about the data generating process. Further research on both the finite sample properties (e.g., unbiasedness or bias, efficiency, mean squared error) and the asymptotic properties (e.g., rates of convergence, limiting distributions) under different assumption of the data generating process would be needed. In this respect, Groeneboom et al. (2001a,b) provide an excellent starting point. The statistical properties of the convex nonparametric quantile regression (CNQR) and the convex asymmetrically weighted least squares (CAWLS) methods introduced in Section 6.4 also deserve further research.

"6. *Investigating the axiomatic foundation of the CNLS and StoNED estimators*."
CNLS regression builds upon the same axioms as DEA, and StoNED estimation applies the minimum extrapolation principle to obtain a unique frontier function that satisfies the postulated axioms. However, it would be compelling if the technology characterized by CNLS and/or StoNED could be stated rigorously from the axiomatic point of view as the intersection of all sets that satisfy the stated axioms and satisfy axiom X. It remains unknown whether axiom X exists, and how it could be formulated explicitly.

"7. *Implementing alternative distributional assumptions and estimating the distribution of the inefficiency term by semi- or nonparametric methods in the cross-sectional setting*."
In this chapter (Section 5.2) we have provided an extensive review of possibilities, including parametric and semi-parametric alternatives. In principle, the quasilikelihood method is applicable to any parametric specification of inefficiency distribution. The most promising way forward seems to be the nonparametric kernel deconvolution of the CNLS residuals, following the works by Hall and Simar (2002) and Horrace and Parmeter (2011). One challenge that remains is to adapt the JLMS conditional mean inefficiency to the semi-parametric setting where no parametric distribution is specified for the inefficiency term.

"8. *Distinguishing time-invariant inefficiency from heterogeneity across firms, and identifying inter-temporal frontier shifts and catching up in panel data models*."
Kuosmanen and Kortelainen (2012) present a simple fixed effects approach to modeling panel data, assuming time-invariant inefficiency. In this chapter we considered the parallel random effects approach, following Eskelinen and Kuosmanen (2013). Ample opportunities for extending these basic techniques to more sophisticated semi-parametric models allowing for technical progress and time-varying inefficiency are available. Indeed, panel data models have been extensively studied both in general econometrics and in the SFA literature. Both the insights and practical solutions from panel data econometrics can be imported to the CNLS and StoNED framework.

"9. *Extending the proposed approach to the estimation of cost, revenue, and profit functions as well as to distance functions*."
Kuosmanen and Kortelainen (2012) consider the estimation of cost function in the single output case under CRS. They made these restrictive assumptions because the cost function must be a concave function of input prices. However, if the standard convexity axiom of the production possibility set holds, then the cost function is a convex function of outputs. A challenge that remains is to formulate the CNLS problem such that we can estimate a function that is

convex in one subset of variables (i.e., outputs), but concave in another subset of variables (i.e., input prices). Kuosmanen (2012) estimates a multi-output cost function using StoNED, but the input prices were excluded by assuming that all firms take the same input prices as given.

"10. *Developing a consistent bootstrap algorithm and/or other statistical inference methods*."
An earlier version of Kuosmanen and Kortelainen (2012) proposed to adapt the parametric bootstrap method proposed by Simar and Wilson (2010) for drawing statistical inferences in the StoNED setting. However, the anonymous reviewers were not convinced that the proposed boostrap method is necessarily consistent when applied to the CNLS residuals. Indeed, one should be wary of naïve bootstrap and resampling approaches that produce invalid and misleading results. Since Kuosmanen and Kortelainen were not able to prove consistency of Simar and Wilson's bootstrap procedure in the CNLS case, the suggestion was excluded from the published version. We stress that adapting one of the known variants of the bootstrap method to the context of CNLS and StoNED would be straightforward. The challenge is to prove that the chosen version of bootstrap method is consistent under the stated assumptions about the data generating process. Another promising approach is to test if CNLS estimates differ significantly from the corresponding estimates obtained using parametric methods (see Sen and Meyer, 2013). As for the contextual variables, Johnson and Kuosmanen (2012) prove that conventional inference techniques from linear regression analysis (e.g., t-tests, p-values, confidence intervals) can be applied for the parametric part (i.e., the coefficients of the contextual variables).

"11. *Conducting further Monte Carlo simulations to examine the performance of the proposed estimators under a wider range of conditions, and comparing the performance with other semi- and nonparametric frontier estimators*."
Several published studies provide Monte Carlo evidence on the finite sample performance of CNLS and StoNED estimators. Kuosmanen (2008) and Kuosmanen and Kortelainen (2012) provide the first simulation results for CNLS and StoNED, respectively, focusing on the precision in estimating the frontier production function f. Johnson and Kuosmanen (2011) present MC simulations regarding the estimation of the parametric δ representing the effect of a single contextual variable z that may be correlated with input x. Andor and Hesse (in press) provide an extensive comparison of the performances of DEA, SFA, and StoNED, mainly focusing on the estimation of the firm specific inefficiency u_i. However, note that all estimators considered are inconsistent in the noisy setting considered because u_i is just a single realization of a random variable. Kuosmanen, Saastamoinen and Sipiläinen (2013) compare performances of DEA, SFA and StoNED in terms of estimating a frontier cost function. They calibrate their simulations to match the empirical characteristics of the Finnish electricity distribution firms. Their simulations demonstrate that

if the premises stated by the Finnish energy regulator hold, then the StoNED estimator has superior performance compared to its restricted special cases, DEA and SFA. As for further research, it would be interesting to compare performance of CNLS and StoNED with those of other semi- and nonparametric frontier estimation techniques such as kernel regression and local maximum likelihood.

"12. *Applying the proposed method to empirical data, and adapting the method to better serve the needs of specific empirical applications*."
The first published application of the StoNED method was Kuosmanen and Kuosmanen (2009), who estimated the production function from the data of 332 Finnish dairy farms in order to assess sustainability performance of farms. Subsequently, there have been several applications in the energy sector, both in production and distribution of electricity. Mekaroonreung and Johnson (2012) applied StoNED to estimate the shadow prices of SO2 and NOx from the data of U.S. coal-fired power plants. Thus far, the most significant real-world application of StoNED has been the study by Kuosmanen (2012) [see also Kuosmanen, Saastamoinen and Sipiläinen (2013) and Dai and Kuosmanen (2014)]. Based on the results of this study, the Finnish energy market regulator adopted the StoNED method in systematic use in the regulation of the Finnish electricity distribution industry, with the total annual turnover of more than €2 Billion. Another real-world application of StoNED is Eskelinen and Kuosmanen (2013), who assessed inter-temporal performance of sales teams using monthly data of Helsinki OP-Pohjola Bank, in close collaboration with the central management of the bank. The results and insights gained in this study were communicated to the team managers and were utilized for setting performance targets for sales teams. These empirical applications illustrate the flexibility and adaptability of the StoNED methodology to suit the specific needs of the application. The applications also provide motivation for developing further methodological extensions to meet the requirements of future applications.

In conclusion, we hope the 12-point program discussed above might inspire future methodological research along the lines described or along new avenues that have escaped our attention. We also hope that the methodological tools currently available would find inroads to empirical applications. In our experience from both Monte Carlo simulations and real empirical applications, CNLS and StoNED has proved dependable, reliable and robust, with an ability to produce results and insights that could not be found using the conventional methods.

References

Afriat, S.N. (1967). The Construction of a Utility Function from Expenditure Data. *International Economic Review* 8: 67-77.

Afriat, S.N. (1972). Efficiency estimation of production functions. *International Economic Review* 13(3): 568-598.

Aigner, D. and S. Chu (1968). On estimating the industry production function. *American Economic Review* 58: 826-839.

Aigner, D., Lovell, C.A.K., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6: 21-37.

Alminidis, P., J. Qian and R. Sickles (2009). Stochastic Frontiers with Bounded Inefficiency, *mimeo*, Rice University.

Almanidis, P. and R. C. Sickles (2011). The skewness issue in stochastic frontier models: Fact of fiction? In I. van Keilegom and P. W. Wilson (Eds.), *Exploring Research Frontiers in Contemporary Statistics and Econometrics*. Springer Verlag, Berlin Heidelberg.

Andor, M. and F. Hesse (in press). The StoNED Age: The Departure Into a New Era of Efficiency Analysis? – A Monte Carlo Comparison of StoNED and the "Oldies" (SFA and DEA). *Journal of Productivity Analysis*, DOI: 10.1007/s11123-013-0354-y

Aragon, Y., A. Daouia, and C. Thomas-Agnan (2005). Nonparametric frontier estimation: A conditional quantile-based approach. *Econometric Theory* 21: 358–389.

Banker, R.D. (1993). Maximum Likelihood, Consistency and Data Envelopment Analysis: A Statistical Foundation. *Management Science* 39, 1265-1273.

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): 1078-1092.

Banker, R.D., and A. Maindiratta (1986). Piece-Wise Loglinear Estimation of Efficient Production Surfaces. *Management Science* 32(1): 126-135.

Banker R.D., and A. Maindiratta (1992). Maximum likelihood estimation of monotone and concave production frontiers. *Journal of Productivity Analysis* 3: 401–415.

Banker, R.D. and R. Natarajan (2008). Evaluating contextual variables affecting productivity using data envelopment analysis. *Operations Research* 56(1): 48-58.

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

Carree, M.A. (2002). Technological inefficiency and the skewness of the error component
in stochastic frontier analysis. *Economics Letters* 77: 101–107.

Caudill, S., and J. Ford (1993). Biases in frontier estimation due to heteroscedasticity. *Economics Letters* 41(1): 17-20.

Caudill, S., J. Ford , and D. Gropper (1995). Frontier estimation and firm-specific inefficiency measures in the presence of heteroscedasticity. *Journal of Business and Economic Statistics* 13(1): 105-111.

Chambers, R.G., Y.H. Chung, and R. Färe (1996). Benefit and distance functions. *Journal of Economic Theory* 70(2): 407-419.

Chambers R.G., Y. Chung and R. Färe (1998). Profit, distance functions and Nerlovian efficiency. *Journal of Optimization Theory and Applications* 98, 351–364.

Charnes, A., W.W. Cooper, and E. Rhodes (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research* 2: 429-444.

Charnes, A., W.W. Cooper, L. Seiford, and J. Stutz (1982). A multiplicative model for efficiency analysis. *Socio-Economic Planning Sciences* 16: 223–224.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In: J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Vol. 6*. North Holland.

Cobb, C.W. and P.H. Douglas (1928). A Theory of Production. *American Economic Review* 18: 139-165.

Coelli, T. (1995), Estimators and hypothesis tests for a stochastic frontier function: A Monte Carlo analysis. *Journal of Productivity Analysis* 6, 247-268.

Coelli, T., and S. Perelman (1999). A comparison of parametric and non-parametric distance functions: With application to European railways. *European Journal of Operational Research* 117(2): 326-339.

Coelli, T., and S. Perelman (2000). Technical efficiency of European railways: a distance function approach. *Applied Economics* 32(15): 1967-1976.

D'Agostino, R., and E.S. Pearson (1973). Tests for Departure from Normality. Empirical Results for the Distributions of b2 and $\sqrt{b_1}$. *Biometrika* 60(3): 613-622.

Dai, X. and T. Kuosmanen (2014). Best-practice benchmarking using clustering methods: Application to energy regulation. *Omega* 42(1): 179-188.

Dantzig, G.B., D.R. Fulkerson, and S.M. Johnson (1954). Solution of a large-scale traveling salesman problem. *Operations Research* 2, 393–410.

Dantzig, G.B., D.R. Fulkerson, and S.M. Johnson (1959). On a linear-programming combinatorial approach to the traveling-salesman problem. *Operations Research* 7, 58–66.

Daouia, A., and L. Simar (2007). Nonparametric efficiency analysis: A multivariate conditional quantile approach. *Journal of Econometrics* 140: 375–400.

Eskelinen, J. and T. Kuosmanen (2013). Intertemporal efficiency analysis of sales teams of a bank: Stochastic semi-nonparametric approach. *Journal of Banking and Finance* 37(12): 5163-5175.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19:1257–1272

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

Farrell, M.J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society Series A* 120: 253-281.

Färe, R., S. Grosskopf, D.-W. Noh, and W. Weber (2005). Characteristics of a polluting technology: theory and practice. *Journal of Econometrics* 126: 469-492.

Färe, R., S. Grosskopf, M. Norris and Z. Zhang (1994). Productivity growth, technical progress, and efficiency change in industrialized countries. *American Economic Review* 84(1): 66–83.

Gabrielsen, A. (1975). On estimating efficient production functions. Working Paper No. A-85, Chr. Michelsen Institute, Department of Humanities and Social Sciences, Bergen, Norway.

Greene, W.H. (1980). Maximum likelihood estimation of econometric frontier functions. *Journal of Econometrics* 13: 26-57.

Greene, W.H. (2008) The econometric approach to efficiency analysis. In H.O. Fried, C.A.K. Lovell, and S.S. Schmidt (Eds.), *The Measurement of Productive Efficiency and Productivity Growth* (pp. 92-250). New York, Oxford University Press Inc.

Groeneboom, P., G. Jongbloed, and J.A. Wellner (2001a). A canonical process for estimation of convex functions: the ‘‘Invelope’’ of integrated brownian motion +t4. *Annals of Statistics* 29:1620–1652.

Groeneboom, P., G. Jongbloed, and J.A. Wellner (2001b). Estimation of a convex function: characterizations and asymptotic theory. *Annals of Statistics* 29:1653–1698.

Gstach, D. (1998). Another approach to data envelopment analysis in noisy environments: DEA+. *Journal of Productivity Analysis* 9(2): 161-176.

Hadri, K. (1999). Estimation of a doubly heteroscedastic stochastic frontier cost function. *Journal of Business and Economic Statistics* 17 (3), 359-363.

Hall, P., and L. Simar (2002). Estimating a changepoint, boundary, or frontier in the presence of observation error. *Journal of the American Statistical Association* 97: 523-534.

Hannah, L.A. and D.B. Dunson (2013). Multivariate convex regression with adaptive partitioning. *Journal of Machine Learning Research* 14: 3207–3240.

Hanson, D.L. and G. Pledger (1976). Consistency in concave regression. *Annals of Statistics* 4(6): 1038-1050.

Harvey, A.C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica* 44(3): 461-465.

Hildreth, C. (1954). Point estimates of ordinates of concave functions. *Journal of the American Statistical Association* 49: 598-619.

Hoff, A. (2007). Second stage DEA: Comparison of approaches for modeling the DEA score. *European Journal of Operational Research*, 181, 425-435.

Horrace, W., and C. Parmeter (2011). Semiparametric deconvolution with unknown error variance. *Journal of Productivity Analysis* 35(2): 129-141.

Johnson, A.L., and T. Kuosmanen (2011). One-stage estimation of the effects of operational conditions and practices on productive performance: asymptotically normal and efficient, root-n consistent StoNEZD method. *Journal of Productivity Analysis* 36 (2), 219-230.

Johnson, A.L., and T. Kuosmanen (2012). One-stage and two-stage DEA estimation of the effects of contextual variables. *European Journal of Operational Research* 220, 559-570.

Jondrow, J., C.A.K. Lovell, I.S. Materov, and P. Schmidt (1982). On estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics* 19, 233-238.

Keshvari, A. and T. Kuosmanen (2013). Stochastic non-convex envelopment of data: Applying isotonic regression to frontier estimation. *European Journal of Operational Research* 231, 481-491.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R., and G.W. Bassett (1978). Regression quantiles. *Econometrica* 46(1): 33–50.

Koenker, R., and G.W. Bassett (1982). Robust tests for heteroscedasticity based on regression quantiles, *Econometrica* 50: 43-61.

Krugman, P. (1992). *The age of diminished expectations: US economic policy in the 1980s*, MIT Press, Cambridge.

Kumbhakar, S.C., S. Ghosh, and J.T. McGuckin (1991). A generalized production frontier approach for estimating determinants of inefficiency in U.S. dairy farms. *Journal of Business and Economic Statistics* 9(3): 279-286.

Kumbhakar, S.C., and C.A.K. Lovell (2000). *Stochastic frontier analysis*. New York, USA, Cambridge University Press.

Kuosmanen, T. (2006): Stochastic nonparametric envelopment of data: Combining virtues of SFA and DEA in a unified framework, MTT Discussion Paper No. 3/2006.

Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. *Econometrics Journal* 11, 308-325.

Kuosmanen, T. (2012). Stochastic semi-nonparametric frontier estimation of electricity distribution networks: Application of the StoNED method in the Finnish regulatory model. *Energy Economics* 34: 2189-2199.

Kuosmanen, T., and M. Fosgerau (2009). Neoclassical versus frontier production models? Testing for the skewness of regression residuals. *Scandinavian Journal of Economics* 111(2): 351-367.

Kuosmanen, T., and N. Kuosmanen (2009). Role of benchmark technology in sustainable value analysis: An application to Finnish dairy farms. *Agricultural and Food Science* 18 (3-4), 302-316.

Kuosmanen, T., and A.L. Johnson (2010). Data envelopment analysis as nonparametric least-squares regression. *Operations Research*, 58, 149-160.

Kuosmanen, T., A.L. Johnson, and C. Parmeter (2013). Orthogonality conditions for identification of joint production technologies: Axiomatic nonparametric approach to the estimation of stochastic distance functions, unpublished working paper (available from the authors by request).

Kuosmanen, T. and M. Kortelainen (2012). Stochastic non-smooth envelopment of data: Semi-parametric frontier estimation subject to shape constraints. *Journal of Productivity Analysis* 38(1), 11-28.

Kuosmanen, T., A. Saastamoinen, and T. Sipiläinen (2013). What is the best practice for benchmark regulation of electricity distribution? Comparison of DEA, SFA and StoNED methods. *Energy Policy* 61: 740-750.

Lee, C-Y, A.L. Johnson, E. Moreno-Centeno, and T. Kuosmanen (2013). A more efficient algorithm for convex nonparametric least squares. *European Journal of Operational Research* 227(2):391-400.

Lim, E., and P.W. Glynn (2012). Consistency of multidimensional convex regression. *Operations Research* 60(1), 196-208.

Lovell, C.A.K., S. Richardson, P. Travers, and L.L. Wood (1994). Resources and functionings: A new view of inequality in Australia (with), in W. Eichhorn, ed., *Models and measurement of welfare and inequality*, pp. 787-807, Springer, Berlin, Heidelberg, New York.

Marschak, J., and W. Andrews (1944). Random simultaneous equations and the theory of production, *Econometrica* 12, 143–205.

McDonald, J. (2009). Using least squares and tobit in second stage DEA efficiency analyses. *European Journal of Operational Research* 197, 792-798.

Meeusen, W., and J. Vandenbroeck (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review* 18(2): 435-445.

Mekaroonreung, M., and A.L. Johnson (2012). Estimating the shadow prices of SO2 and NOx for U.S. coal power plants: A convex nonparametric least squares approach. *Energy Economics* 34(3): 723-732.

Newey, W.K., and J.L. Powell (1987). Asymmetric least squares estimation and testing. *Econometrica* 55(4): 819-847.

Ondrich, J., and J. Ruggiero (2001). Efficiency measurement in the stochastic frontier
model. *European Journal of Operational Research* 129, 434-442.

Ruggiero, J. (2004). Data envelopment analysis with stochastic data. *Journal of the Operational Research Society* 55(9): 1008-1012.

Schmidt, P., and T. Lin, (1984). Simple tests of alternative specifications in stochastic frontier models. *Journal of Econometric*, 24: 349-361.

Schmidt, P., and R.C. Sickles (1984). Production frontiers and panel data. *Journal of Business and Economic Statistics* 2(4): 367-74.

Seijo, E., and B. Sen (2011). Nonparametric least squares estimation of a multivariate convex regression function. *Annals of Statistics*, 39(3): 1633-1657.

Sen, B., and M. Meyer (2013). Testing against a parametric regression function using ideas from shape restricted estimation, arXiv preprint arXiv:1311.6849, available at: http://arxiv.org/pdf/1311.6849.pdf.

Simar, L., and P.W. Wilson (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science* 44(1): 49-61.

Simar, L., and P.W. Wilson (2000). A general methodology for bootstrapping in non-parametric frontier models. *Journal of Applied Statistics* 27(6): 779-802.

Simar, L., and P.W. Wilson (2007). Estimation and inference in two-stage, semi-parametric models of production processes, *Journal of Econometrics* 136(1): 31-64.

Simar, L., and P.W. Wilson (2010). Inferences from cross-sectional, stochastic frontier models. *Econometric Reviews* 29(1): 62-98.

Simar, L., and P.W. Wilson (2011). Two-stage DEA: Caveat emptor. *Journal of Productivity Analysis* 36(2): 205-218.

Timmer, C.P. (1971). Using a probabilistic frontier production function to measure technical efficiency. *Journal of Political Economy* 79: 767-794.

Varian, H.R. (1984). The nonparametric approach to production analysis. *Econometrica* 52, 579-598.

Verbeek, M. (2008). *A Guide to Modern Econometrics*. England, John Wiley & Sons Ltd.

Wang, H., and P. Schmidt (2002). One step and two step estimation of the effects of exogenous variables on technical efficiency levels. *Journal of Productivity Analysis* 18, 129-144.

Wang, Y., S. Wang, C. Dang, and W. Ge (2014). Nonparametric quantile frontier estimation under shape restriction. *European Journal of Operational Research* 232: 671-678.

Winsten, C.B. (1957). Discussion on Mr. Farrell's Paper. *Journal of the Royal Statistical Society Series A* 120(3): 282-284.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48(4): 817–838.

Article 2

Timo Kuosmanen; Antti Saastamoinen; Timo Sipiläinen. What is the best practice for benchmark regulation of electricity distribution? Comparison of DEA, SFA and StoNED methods. Energy Policy, 2013, vol. 61, pp. 740-750.
DOI: 10.1016/j.enpol.2013.05.091

Energy Policy 61 (2013) 740–750

Contents lists available at ScienceDirect

Energy Policy

journal homepage: www.elsevier.com/locate/enpol

What is the best practice for benchmark regulation of electricity distribution? Comparison of DEA, SFA and StoNED methods

Timo Kuosmanen [a,*], Antti Saastamoinen [a], Timo Sipiläinen [b]

[a] School of Business, Aalto University, P.O. Box 21220, 00076 Aalto, Helsinki, Finland
[b] Department of Economics and Management, University of Helsinki, P.O. Box 27, 00014 Helsinki, Finland

HIGHLIGHTS

- We compare DEA, SFA and StoNED methods in the context of regulation of electricity distribution.
- Both empirical comparisons and Monte Carlo simulations are presented.
- Choice of benchmarking method has a significant economic impact on the regulatory outcomes.
- StoNED yields the most precise results in the Monte Carlo simulations.
- Five lessons concerning heterogeneity, noise, frontier, simulations, and implementation.

ARTICLE INFO

Article history:
Received 29 May 2012
Accepted 24 May 2013
Available online 20 June 2013

Keywords:
Frontier estimation
Nonparametric production analysis
Productive efficiency

ABSTRACT

Electricity distribution is a natural local monopoly. In many countries, the regulators of this sector apply frontier methods such as data envelopment analysis (DEA) or stochastic frontier analysis (SFA) to estimate the efficient cost of operation. In Finland, a new StoNED method was adopted in 2012. This paper compares DEA, SFA and StoNED in the context of regulating electricity distribution. Using data from Finland, we compare the impacts of methodological choices on cost efficiency estimates and acceptable cost. While the efficiency estimates are highly correlated, the cost targets reveal major differences. In addition, we examine performance of the methods by Monte Carlo simulations. We calibrate the data generation process (DGP) to closely match the empirical data and the model specification of the regulator. We find that the StoNED estimator yields a root mean squared error (RMSE) of 4% with the sample size 100. Precision improves as the sample size increases. The DEA estimator yields an RMSE of approximately 10%, but performance deteriorates as the sample size increases. The SFA estimator has an RMSE of 144%. The poor performance of SFA is due to the wrong functional form and multicollinearity.

1. Introduction

Electricity distribution firms typically enjoy a natural local monopoly. This creates a need to regulate the distribution sector. In the theory of regulation, it is well known that the 'cost-of-service' type of pricing does not provide incentives for the electricity distribution firms to minimize the cost (Laffont and Tirole, 1993). To determine a more objective yardstick for the acceptable cost level, Shleifer (1985) suggested comparing the observed cost of a firm with that of its competitors. However, as Pollit (2005) points out, it is often difficult to find exactly identical or even sufficiently similar competitors that could serve as an appropriate yardstick. Instead of using a discrete set of benchmark firms, one could apply frontier estimation methods to estimate a continuous frontier cost function that represents the best practice benchmark. Benchmark regulation has been applied as an integral part of the regulatory framework in many countries (Jamasb and Pollit, 2001; Jamasb et al., 2003, 2004). According to the recent study by Bogetoft and Otto (2011, Ch. 10), at least nine European regulators currently apply the axiomatic DEA (*data envelopment analysis*; Charnes et al., 1978; Farrell, 1957) and the econometric SFA (*stochastic frontier analysis*; Aigner et al., 1977), or some combination thereof.

Ever since the DEA and SFA approaches have been introduced to regulation, there has been lively debate about the suitability of these methods for the purposes of regulation (Dassler et al., 2006; Irastorza, 2003). There is large and growing academic literature on

* Corresponding author. Tel.: +358 9 43131; fax: +358 943 13 8700.
E-mail addresses: timo.kuosmanen@aalto.fi (T. Kuosmanen), antti.saastamoinen@aalto.fi (A. Saastamoinen), timo.sipilainen@helsinki.fi (T. Sipiläinen).

0301-4215/$ - see front matter © 2013 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.enpol.2013.05.091

the application of DEA and SFA in the electricity distribution industry (Agrell et al., 2005; Cullmann, 2009, 2012; Forsund and Kittelsen, 1998; Hjalmarsson and Veiderpass, 1992; Iglesias et al. 2010; Jamasb and Pollit, 2003; Kopsakangas-Savolainen and Svento, 2008; Korhonen and Syrjänen, 2003; Weyman-Jones, 1991). As yet, however, there is no clear conclusion on which method is superior. The inconclusive results have raised concerns about the suitability of any single method for the purposes of benchmark regulation. Thus, many regulators have recently opted to use a combination of both DEA and SFA (see also Azadeh et al. (2009)). In Germany, for example, the regulator estimates efficiency of each firm using both DEA and SFA, and then chooses the larger of the two estimates (Agrell and Bogetoft, 2007). According to Bogetoft and Otto (2011; Ch. 10), at least four European regulators apply some combination of both DEA and SFA.

The Finnish Energy Market Authority (*Energiamarkkinavirasto*, EMV) is one of the pioneers in the practical implementation of benchmark regulation. EMV has used frontier methods as an integral part of the regulatory model since 2005, starting with DEA (Korhonen and Syrjänen, 2003), adopting SFA in 2008 (Syrjänen et al., 2006). In 2010, EMV commissioned several studies to address the critique of DEA and SFA presented by the distribution firms and the energy industry. After a rigorous evaluation process, EMV considered the report by Kuosmanen et al. (2010, in Finnish) as the most promising attempt to overcome the pitfalls of DEA and SFA. Following the recommendation of that report, in 2012 EMV replaced DEA and SFA by the new StoNED method (*stochastic semi-nonparametric envelopment of data*) introduced by Kuosmanen and Kortelainen (2012).

The purpose of this paper is to present a systematic comparison of the DEA, SFA, and a recently proposed StoNED method in the context of energy regulation. Focusing on the model specifications actually employed by EMV during the third regulation period in 2012–2015, we compare the efficiency estimates produced by these three different methods. We also include the average of DEA and SFA efficiency scores to our comparison. This approach was applied by EMV during the second regulation period, 2008–2011. We label it here as naïve model averaging (NMA). Our focus is to examine the observed differences between the methods and discuss the feasibility of the models in regulatory context. More importantly, we also compare the implications of the methodological choices on the monetary cost targets. While the efficiency scores obtained with different methods are usually highly correlated, the economic implications in terms of the cost targets are substantial.

The empirical comparisons show that the choice of the benchmarking method matters in practice. However, empirical comparisons do not allow us to conclude that one method is better than another. Therefore, we also compare the precision of the estimators in the controlled environment of Monte Carlo simulations. A novel feature of our simulations is that we calibrate the data generation process of the simulations to match the essential characteristics of the EMV data as closely as possible to ensure the relevance of the simulation evidence for the real-world regulation. The customized data generation process of the simulations enables us to measure performance of the alternative estimators in the specific context of the EMV's regulatory framework. Our simulation evidence shows that the StoNED method outperforms the conventional DEA, SFA, and their average at all sample sizes considered.

The rest of the paper is organized as follows. Section 2 briefly describes the regulation of Finnish electricity distribution firms and the empirical data used in this study. Section 3 briefly introduces the benchmarking methods considered in this study (more detailed presentation of the methods is available in the online Supplement), and compares the empirical cost frontiers that the methods produce using the data and model specifications of EMV. Section 4 presents an empirical comparison of the efficiency estimates produced by the alternative methods. In Section 5 we briefly comment the implementation of the methods in the EMV regulatory model. Section 6 presents a systematic comparison of the methods in the controlled environment of Monte Carlo simulations. Section 7 summarizes the lessons learned from this study. Additional materials, including a technical appendix that provides a more detailed description of the methods considered, and the computer program used in the simulations, are available online as supplementary materials to this article (see http://www.sciencedirect.com).

2. Benchmark regulation of electricity distribution in Finland

In the regulation of the Finnish electricity distribution firms, EMV applies a combination of the traditional revenue cap and the benchmarking regimes. In the EMV model, all distribution firms are systematically assessed every year. The annual revenue figures of each firm are compared with the acceptable level of revenue to calculate the annual surplus or deficit. The acceptable revenue figure includes the acceptable total costs plus the acceptable rate of return for the invested capital, which is calculated based on the capital asset pricing model. As a part of determining the acceptable total cost, EMV applies the cost frontier model as a benchmark, as will be discussed in more detail below. At the end of the four-year regulation period, EMV calculates the total surplus or deficit accumulated over the regulation period, which needs to be balanced during the next 4-year period. A firm can return the surplus to the customers by charging lower tariffs in the next regulation period, whereas the deficit allows a firm to increase its tariffs in the next regulation periods. Kinnunen (2006) provides a more detailed review of the EMV model from the perspective of investment incentives. Kuosmanen (2012) discusses the recent reforms in the benchmark regulation and the incentives for improving productivity and efficiency. Further information about the Finnish regulatory model can be found on the EMV website: http://www.emvi.fi.

In the current regulation period, in years 2012–2015, the regulation of the acceptable total cost is based on the following generic cost frontier:

$$\ln x = \ln C(y_1, y_2, y_3) + \delta z + u + v \tag{1}$$

where

x is the observed total cost (TOTEX) (€1000),
C is the frontier cost function,
y_1 is the energy transmission (GWh),
y_2 is the total length of the network (km),
y_3 is the number of customers,
z is the proportion of underground cables,
δ is the coefficient of the z variable,
u is the random variable representing inefficiency, and
v is the random variable representing stochastic noise.

In this study the cost variable x refers to the total expenditure (TOTEX), which consists of three components: controllable operational costs (OPEX), capital expenditures (CAPEX) and the external supply interruption costs for customers (INT).[1] The last component

[1] Our empirical comparison is based on the original data and the model specification recommended in Kuosmanen et al. (2010) and Kuosmanen (2012). EMV has made some subsequent modifications to the model and the data. In the model implemented by EMV, the observed annual capital expenditures (CAPEX) are included in the acceptable total cost as such, and the benchmark regulation is only

Table 1
Descriptive statistics of variables.

Variable	Mean	St. dev.	Min.	Max.
x=Total cost (€1000)	8418.91	18,047.78	267.81	117,554.10
y_1=Energy transmission (GWh)	480.39	971.51	14.81	6599.71
y_2=Length of network (km)	4135.27	10,223.27	50.80	67,611.05
y_3=No. of customers	35,448.68	71,870.65	24.25	420,473.00
z=Proportion of underground cables	0.33	0.26	0.01	1.00

can also be viewed as a quality component, as the lack of supply interventions can be interpreted as an indicator of good service. Since the outputs are almost time-invariant throughout the period, all variables are defined as the yearly averages over the period 2005–2008 (see Kuosmanen (2012), for a discussion). Before averaging, the total costs are deflated to the prices of 2005. In this specification, inefficiency u represents the average inefficiency over the evaluation period. Averaging of data also reduces the variance of the noise term v.

The output variables are the weighted amount of energy transmitted through the network (y_1, GWh, of 0.4 kV equivalents), the total length of the network (y_2, km), and the total number of customers connected to the network (y_3, number). In y_1, the transmission of electricity at different voltage levels is weighted according to the average cost of transmission such that the high-voltage transmission gets a lower weight than the low-voltage transmission. Note that y_1 depends on the observed demand for electricity, whereas outputs y_2 and y_3 capture the potential or latent demand and are thus defined as outputs in the regulatory model (see Kuosmanen (2012) for further details). In essence, outputs y_2 and y_3 capture the fixed cost of maintaining a sufficient capacity to provide service for the given network area irrespective of the actual consumption of electricity.

In addition to the three outputs, the latest EMV specification introduced a contextual variable z, defined here as the proportion of underground cables in the total length of the network. The z-variable is not an input or output as such; it controls the heterogeneity of the firms and their operating environments. Note that the contextual variable enters Model (1) in a parametric form, analogous to the standard regression analysis, while the output variables can be modeled using either a parametric or nonparametric specification of the cost function C. If nonparametric specification of C is assumed, it is then appropriate to characterize Model (1) as a semi-nonparametric, partially linear model of cost frontier. Modeling contextual variables in this way allows us to capture the average effect of underground cabling on cost (represented by the coefficient δ), without increasing the number of explanatory variables included in the nonparametric part which is subject to *the curse of dimensionality* (Simar and Wilson, 2008).

Our data consists of 89 Finnish electricity distribution companies, whose networks cover practically all regions of Finland. Table 1 presents the descriptive statistics for total costs, three outputs, and the underground cabling variable, which describes the operational conditions of a company (see Section 3.1 for details). Recall that our data are four year averages of years 2005–2008.

Table 1 reveals that the industry consists of a very heterogeneous set of firms. For example, the size of companies measured by the amount of transmitted energy varies from 15 to 6600 GWh per year. There are also considerable differences in the operating environments of the firms. On average, the proportion of underground cabling is 33% but the range is almost from 1% to 100%. The proportion of underground cabling is highest in the dense urban areas. Note that the data also includes some industrial network operators, which transmit a large amount of energy to a small number of industrial customers.

(*footnote continued*)
applied to the controllable operational expenditures (OPEX) plus a half (i.e., 50%) of the interruption costs (INT).

3. Comparison of empirical cost frontiers

This section introduces the benchmarking methods considered in this study by comparing the empirical estimates of the cost frontiers obtained by each method. We believe the empirical frontier estimates aptly illustrate the information content and the comparative advantages of the methods considered. Readers interested in the technical details of the methods can consult the technical appendix provided as an online Supplement, or the references provided below. The StoNED method is presented in detail in Kuosmanen and Kortelainen (2012) and Kuosmanen (2012). Detailed presentations of the conventional DEA and SFA are available in numerous articles and books (Fried et al., 2008).

3.1. StoNED frontier

We start the empirical comparison with the StoNED method, which EMV adopted for the current regulation period in years 2012–2015. The main appeal of StoNED is its ability to accommodate the main advantages of both DEA and SFA: it combines the non-parametric, piece-wise linear DEA-style frontier with the stochastic SFA-style treatment of inefficiency and noise. This makes StoNED more robust to both model misspecification and noise.[2] A detailed presentation of the model specification applied by EMV can be found in Kuosmanen (2012). Therefore, we will here discuss only some general properties of the method.

StoNED does not require any *a priori* assumptions about the functional form of the cost frontier. Similar to DEA, StoNED imposes general axioms concerning the benchmark technology, such as monotonicity, convexity, and returns-to scale.[3] Throughout this study we assume constant returns to scale to hold for distribution companies (see Kuosmanen (2012), for details). On the other hand, StoNED model incorporates the core aspects of SFA by including both inefficiency and noise as possible sources of deviation from a benchmark technology. Kuosmanen and Kortelainen (2012) operationalize the StoNED model by formulating it as a convex nonparametric least squares (CNLS) problem.

[2] Previous published applications of the StoNED method are in the areas of agriculture (Kuosmanen and Kuosmanen, 2009), electricity generation (Mekaroonreung and Johnson, 2012), electricity distribution (Kuosmanen, 2012), and banking (Eskelinen and Kuosmanen, in press).

[3] The term benchmark technology refers to the frontier used as a point of reference in productivity and efficiency assessment. The axioms of the benchmark technology represent our ex ante requirements for efficient performance (e.g., monotonicity stems from the definition of technical efficiency by Koopmans, 1951). The underlying production technology does not necessarily need to satisfy all these axioms.

Table 2
StoNED marginal costs and average efficiencies by firm groups (CRS).

Group	No. of firms	Energy transmission (€ cents/kWh)	Network length (€/km)	No. of customers (€/customer)	Average efficiency (%)
1	11	0.6043	876.74	0.87	92
2	36	0.5597	984.94	1.23	92
3	10	0.4566	1038.81	1.86	93
4	3	0.4434	908.77	22.25	94
5	3	0.4200	970.69	21.00	92
6	4	0.3662	964.71	27.86	95
8	7	0.3493	930.93	33.43	91
9	6	0.3324	983.05	29.61	90
7	3	0.2929	232.21	60.11	92
Others	6				96
Average		0.4773	930.09	12.94	92

Contextual variables z were introduced by Johnson and Kuosmanen (2011, 2012). Note that while the frontier itself is specified in a fully nonparametric fashion, in the second stage the inefficiency and noise terms are distinguished by means of some distributional assumption. Thus, it is appropriate to classify the method as semi-nonparametric.

The piecewise linear of frontier that StoNED produces allows the marginal costs differ between the firms. In other words, the reference unit for the firms may be located at different segments of the frontier. The linear segments that constitute the frontier have different slopes. This offers more flexibility in terms of technology and in addition of the z-variable it partly accounts for the heterogeneity of the firms.

Table 2 presents the firm-specific estimates of the average marginal costs for 10 groups of firms, grouped according to the estimated beta coefficients.[4] As in DEA, the standard errors for these coefficient estimates are not readily available. The groups have been sorted in a descending order according to the marginal cost on energy transmission. These marginal costs are the most favorable ones for each company: no company could increase its efficiency by deviating from the marginal costs implied by StoNED even if the regulator allowed firms to freely choose their marginal costs.

The average marginal costs are reported in the bottom row of Table 2. The estimated marginal costs (0.48 c/kWh for electricity transmission, 930 €/km for network length, and 13 €/user) appear reasonable based on our experience of this sector (cf., Tables 3 and 4). Firm-specific coefficients can differ substantially from these average values. For example, the marginal cost per user is lowest in Group 1, which consists of firms operating in rural areas, whereas the marginal cost per user is highest in Group 7, consisting of city firms. The last column of Table 2 reports the average cost efficiency (CE) of firms within each group. While there are differences in marginal costs, the differences in the average cost efficiency levels are relatively small. This suggests that the method does not systematically favor some firms over others due to their operational environment.

3.2. DEA frontier

EMV applied DEA in the first two regulation periods in 2005–2007 and 2008–2011 (see Korhonen and Syrjänen (2003), for further discussion). Similar to StoNED, DEA is an axiomatic, nonparametric approach to estimate the frontier. In fact, Kuosmanen and Johnson (2010) have shown that DEA can be obtained as a restricted special case of CNLS formulation of the StoNED model. Both the methods are based on the same set of axioms. The only notable difference between the methods is their assumption about the deviations from the frontier. Whereas StoNED assumes the deviations to consist from two elements, inefficiency and noise, DEA assumes only inefficiency. This is generally seen as the main shortcoming of DEA. DEA is also sensitive to outlier observations as it fully envelops the data based on the outermost observation in each dimension. In other words, often only few observations determine the frontier.

[4] See Kuosmanen (2012) for a 3-dimensional graphical illustration of the estimated StoNED frontier.

Table 3
DEA marginal costs by firm groups (CRS).

Group	Number of firms	Energy transmission (€ cents/kWh)	Network length (€/km)	No. of customers (€/customer)	Average efficiency (%)
1	2	0.5972	0	54.46	100
2	8	0.5910	489.33	41.28	79
3	23	0.5866	846.33	15.42	77
4	3	0.5857	930.62	0	65
5	12	0.5494	958.45	0	80
6	2	0.3604	494.77	71.72	85
7	7	0.3504	863.33	45.77	84
8	3	0.1491	1142.03	0	80
9	3	0	0	133.71	84
10	2	0	182.59	128.49	85
11	5	0	606.91	111.63	76
12	16	0	820.00	95.85	83
13	3	0	1069.20	34.67	85
Average		0.3526	762.47	46.20	80

Table 4
Marginal costs of outputs estimated by SFA; in Model A the total network length is used; in Model B the urban network (y_{2A}) and other network (y_{2B}) are treated as separate outputs.

	Model A	Model B
y_1: Energy trans. (€ cents/kWh)	0.61** (0.000)	0.60** (0.000)
y_2: Network length (€/km)	896.74** (0.000)	–
y_{2A}: Urban network (€/km)	–	1115.94** (0.001)
y_{2B}: Other network (€/km)	–	904.06** (0.000)
y_3: No. of customers (€/customer)	25.32 (0.114)	20.12 (0.264)

p-Values in parenthesis.
Statistical significance indicated as follows: * refers to 5% significance, ** refers to 1% significance.

The marginal costs (shadow prices) of outputs estimated by DEA are presented in Table 3. Analogous to Table 2, firms have been classified to 13 groups in a descending order with respect to the marginal cost on energy transmission. The figures are the average marginal cost in each of the groups. Note that for many groups the marginal cost equals zero. In particular, the estimated marginal cost of energy transmission is zero for five groups (29 firms). This can partly explain why the average of the DEA estimates for the marginal cost of energy transmission (0.35 c/kWh) is lower than the corresponding StoNED estimate (0.48 c/kWh).

Recall that the DEA frontier envelops all observations, attributing all deviations from frontier to inefficiency, whereas the StoNED frontier takes the noise explicitly into account. Therefore, we can expect that the DEA estimates of firm-specific marginal costs are generally lower than the corresponding StoNED estimates since

DEA envelops all data whereas noise is included in StoNED. The average DEA shadow price is indeed lower than the StoNED estimate for the network length (DEA: 762 €/km; StoNED: 930 €/km). As for the marginal cost per user, the DEA estimate is notably higher than the StoNED estimate (DEA: 46 €/user; StoNED: 13 €/user). This implies that the shapes of the estimated DEA and StoNED cost frontiers differ considerably, particularly for the output profile of the urban networks that assign a high shadow price for the number of customers (Groups 9–12 in Table 3).

The rightmost column of Table 3 reports the average efficiency of firms projected to each facet of the DEA frontier. We note that the differences in average efficiencies across facets are notably larger in DEA than in StoNED (compare with Table 2).

3.3. SFA frontier

EMV applied SFA and DEA in parallel during the second regulation period in 2008–2011. In practice, EMV applied the unweighted average of the DEA and SFA efficiency scores to set cost reduction targets for each firm (referred to as NMA in this paper).

The econometric SFA approach requires some parametric assumptions concerning the functional form of the cost frontier. The Cobb–Douglas and translog are the most commonly used functional forms applied in the SFA literature. The SFA model can be obtained as a special case of the generic cost frontier Model (1), obtained by imposing some specific functional form for the cost function C. In fact, SFA results as a restricted as a special case of StoNED if the functional form is restricted to be linear and it is assumed that marginal costs are equal for all firms.[5]

The main shortcoming of SFA is that the functional form assumptions are somewhat arbitrary and difficult to justify. In the present context, most commonly used functional forms fail to capture the economies of scope in joint production (Syrjänen et al., 2006). For example, the standard Cobb–Douglas function is quasi-concave at all parameter values. This implies the Cobb–Douglas cost function exhibits economies of specialization rather than economies of scope which again could give wrong incentives to specialize in provision of just one output instead of a balanced portfolio of outputs. The flexible functional forms such as translog are subject to the same problem, and the larger number of parameters would likely cause additional problems with multicollinearity. This is why EMV used the linear functional form for SFA in the previous regulation period. Linear functional form however assumes that outputs are perfect substitutes. and thus it tends to favor the "average firm" over the firms operating with an atypical output profile.

As was apparent from the general regulatory model presented in Eq. (1), the heterogeneity of the firms must also be taking account. As a partial adjustment to the heterogeneity of firms and their operating environments, the total network length y_2 was divided in two parts in the SFA model EMV applied in the previous regulation period 2008–2011), specifically,

$$y_2 = y_{2A} + y_{2B}, \tag{2}$$

where

y_{2A}=length of underground cabled urban network (km), and
y_{2B}=length of other network (km).

Treating y_{2A} and y_{2B} as separate outputs in the SFA model, the marginal cost of the underground cabled urban network is allowed to be higher than that of the other network. This however is slightly problematic from the point of view EMV averaging approach as now the components of the average are based on different model specification (see details in the technical appendix provided in the Supplement).

The SFA estimates of the marginal costs of outputs are presented in Table 4. For completeness, we report the estimates for the three-output model where the total network length (y_2, Model A) is used as an output and for the four-output model where the network length is separated in two components (y_{2A} and y_{2B}, Model B). The SFA model is estimated by maximum likelihood assuming CRS (i.e. the intercept term has been set to zero). In case of SFA, here and in Section 4, we assume the truncated normal distribution for the inefficiency distribution, as this is the specification that EMV used in the previous regulation period, following Syrjänen et al. (2006).

Comparing the results of Table 4 with the marginal costs reported in Tables 2 and 3, we find that the marginal costs suggested by SFA differ from the average marginal costs estimated by DEA or StoNED. For energy transmission, for example, the marginal cost estimates obtained by SFA are notably larger than the average of the StoNED estimates (only for Group 1 in Table 2, the marginal cost is close to the SFA estimates), and almost twice as large as the average of DEA estimates (Groups 1–4 in Table 3 yield marginal costs nearly as high as the SFA estimates).

In Model B, the estimated marginal cost of underground cabled urban network is higher than that of the other network, as expected. Note that the marginal cost of the total network length in Model A is lower than the marginal cost of the other network in Model B. Division of the network length on two parts has little effect on the marginal cost of the energy transmission, but does have a notable impact on the marginal cost per user. Clearly, taking the heterogeneity of firms into account influences the marginal cost estimates. Recall that the nonparametric DEA and StoNED methods allow for firm-specific marginal costs, which provides greater flexibility in terms of the heterogeneity of firms and their operating environments, as discussed at the end of Section 3.1.

The SFA estimate for the marginal cost per user is relatively small and insignificant at the conventional significance levels. The StoNED estimates for the marginal cost per user are larger for some groups (particularly firms operating in large cities), but the average of StoNED estimates falls below the SFA estimate. The DEA estimates are notably larger, for three groups the marginal cost estimate exceeds €100 per user. For firms operating in rural areas, the number of customers is not the main cost driver; majority of Finnish distribution networks operate in rural areas. This explains why the SFA estimate and the averages of DEA and StoNED estimates of the marginal cost per user are rather low.

The SFA results reported in Table 4 have been estimated using the heteroskedasticity correction following Syrjänen et al. (2006), who assume that the variances of the inefficiency and noise terms are proportional to the amount of transmitted energy (y_1). It is likely that the deviations from the cost frontier are dependent from the company size. In econometrics, the textbook treatment of such heteroskedasticity is to normalize all variables by y_1. The assumed form of heteroskedasticity however appears completely arbitrary: one could equally well assume that heteroskedasticity is driven by any other output of combination of them. To examine the effect of heteroskedasticity correction in more detail, we have estimated the SFA model again using each output variable as the normalizing criterion, and without any normalization. The SFA models are estimated with modified OLS (MOLS) and the parameter estimates of the models with alternative normalizations are reported in Table 5, both under CRS (the top part) and variable returns to scale (VRS, the bottom part).[6]

[5] The random parameters SFA models (Tsionas, 2002; Greene, 2005) allow for heterogeneity across firms by introducing firm-specific coefficients.

[6] The maximum likelihood estimator of the SFA model fails due to wrong skewness of residuals in six out of the eight specifications considered. Thus, for this

Table 5
Impact of normalization on SFA (MOLS estimates).

	Normalization			
	None	By energy	By other network	By customers
CRS model				
Energy (€ cents/kWh)	0.09	1.00**	0.49**	0.57**
Urban network (€/km)	−462.17	1464.38**	−2445.20**	16,340.92**
Other network (€/km)	1044.65**	916.59**	1367.96**	248.66
Users (€/user)	113.03**	−0.67	56.00**	−62.62
Expected efficiency (%)	#	72	§	#
R^2	0.998	0.876	0.949	0.997
VRS model				
Constant (t€)	108.85	111.82**	−183.51**	713.27**
Energy (€ cents/kWh)	0.08	0.90**	0.56**	0.44**
Urban network (€/km)	−534.79	1067.59**	−994.90	−279.20
Other network (€/km)	1053.56**	854.50**	1412.23**	578.94**
Users (€/user)	113.82**	10.50	50.39**	40.49
Expected efficiency (%)	#	81	§	#
R^2	0.997	0.899	0.961	1.000

Statistical significance indicated as follows: * refers to 5% significance, ** refers to 1% significance.
# Indicates negative skewness (negative $\hat{\sigma}_u$).
§ Indicates too large a skewness (negative $\hat{\sigma}_v$).

Table 5 shows that the choice of the normalization has a major impact on the marginal costs of outputs and some results do not seem to be very meaningful. Many marginal costs are negative; only in the VRS model normalized by energy transmission all coefficients are positive as expected. The normalization also influences the skewness of the residuals. If no normalization is applied, or the normalization is based on the number of customers, then the skewness of the OLS residuals has a wrong sign, and hence the stochastic frontier reduces to the OLS curve (Kumbhakar and Lovell, 2000). In Table 5, these cases are indicated by # on the row "Expected efficiency". On the other hand, if the normalization is based on the network length, the skewness is so large that the estimate of $\hat{\sigma}_v$ becomes negative. These cases are indicated by §. Thus, we find that the normalization by energy transmission is not only important for heteroskedasticity correction: it is the only specification in Table 5 that yields meaningful efficiency estimates as well as positive marginal costs in the VRS case. We suspect the parameter estimates are sensitive to the choice of normalization due to multicollinearity of output variables.

4. Comparison of efficiency estimates

The previous section presented some selected empirical evidence of the cost frontier obtained with different methods. In this section we compare the empirical estimates of cost efficiency (CE).[7] Our focus on the CE scores is motivated by the fact that the Finnish legislation mandates the use the efficiency improvement targets as the regulatory instrument of EMV. We also include the efficiency scores obtained through *naïve model averaging* (NMA) to our comparison. These figures are simply the averages of the DEA and SFA estimates. The practical justification of NMA was to alleviate the possible modeling misspecification of SFA and DEA by taking an average of the two efficiency estimates.[8]

(*footnote continued*)
comparison, we report the Modified OLS (Aigner et al., 1977; Olson et al., 1980) estimates throughout all eight specifications considered in Table 5.

[7] For all methods, we follow the model specifications applied by EMV. For comparability, CRS is imposed throughout all estimation methods considered.

Table 6
Correlation analysis of efficiency scores.

	Pearson correlation				Spearman rank-correlation			
	StoNED	DEA	SFA	NMA	StoNED	DEA	SFA	NMA
StoNED	1	0.9089	0.8956	0.9367	1	0.9338	0.8788	0.9498
DEA		1	0.8568	0.9726		1	0.8456	0.9732
SFA			1	0.9523			1	0.9329
NMA				1				1

Table 7
Descriptive statistic of efficiency scores.

	Mean	St. Dev.	Median	Min.	Max.
StoNED	0.924	0.069	0.940	0.764	1.000
DEA	0.802	0.119	0.807	0.466	1.000
SFA	0.862	0.092	0.892	0.545	0.981
NMA	0.832	0.102	0.848	0.505	0.990

Consider first the correlations between the CE scores estimated by the four methods. Table 6 reports the correlation matrices of the Pearson product moment correlation coefficients (the left side), and the Spearman rank correlation coefficients (the right side). There is a high positive correlation in every pair of CE estimates. Based on the correlation analysis alone, one might be tempted to conclude the choice of the estimation method has little effect on the efficiency estimates. However, this conclusion proves wrong in a closer inspection of the levels of CE estimates.

Table 7 reports descriptive statistics of the *CE* scores obtained by different methods. There are notable differences in the levels of efficiency scores. In particular, we find that StoNED yields considerably higher efficiency scores than any other method, both in terms of the mean and the minimum: recall that StoNED takes the noise term explicitly into account and captures heterogeneity of firms and their operating environments through the use of the contextual variable z, which is omitted in other methods. [9]

The summary statistics of Table 7 facilitate the comparisons of an average or a median firm. To shed further light on efficiency of individual firms, we have plotted the StoNED efficiency scores against the NMA estimates in Fig. 1. Points in this diagram represent the pair of efficiency estimates obtained by the average of DEA and SFA (NMA, the horizontal axis) and StoNED (the vertical axis). The broken line in the middle of diagram indicates the 45° line: for points above this line the StoNED efficiency estimate is greater than that of NMA.

Fig. 1 illustrates that the StoNED estimator is more favorable for each individual firm than the average value of DEA and SFA; the StoNED efficiency scores are higher than the corresponding NMA values. For some companies the use of NMA value would yield efficiency improvement targets around 35–50% (efficiency of 50–65%). Improvements of this magnitude seem highly unrealistic. Note that there are many firms that lie relatively close to the

[8] Similar practice of combining DEA and SFA estimators has been used or considered for use in other countries as well, see, e.g., Pollit (2005), Azadeh et al. (2009), and Bogetoft and Otto (2011, Ch. 10).

[9] Note that in StoNED the probability mass at $u=0$ is equal to zero, and hence none of the firms are 100% efficient. Still, the maximum value is rounded to 1.000 at the accuracy of three decimal digits.

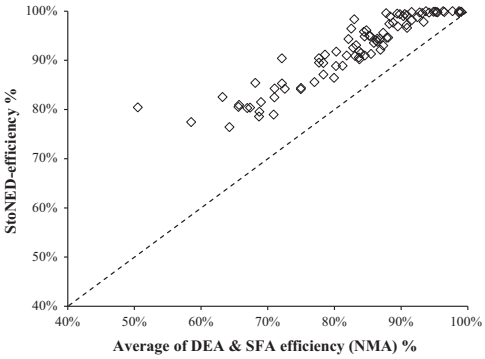

Fig. 1. Comparison of StoNED and NMA efficiency scores.

Table 8
Monetary cost reduction targets (thousand € in prices of 2008).

	Industry	Mean	St. Dev.	Min.	Max.
StoNED	47,508	534	1326	0.000	11,113
DEA	141,382	1589	3888	0.000	27,654
SFA	93,023	1045	2185	0.024	13,599
NMA	117,205	1317	2947	0.017	20,627

efficient StoNED frontier. This is in contrast especially with DEA, where only a few observations define the efficient frontier.

The main objective of the efficiency estimates is to provide cost targets that EMV imposes to the distribution firms. To examine the impacts of the methodological choice on the bottom line, we have converted the firm-specific efficiency estimates to monetary cost reduction targets, calculated as $x_i(1-CE_i)$, where CE_i is the firm-specific estimate of inefficiency and x_i is the observed total cost in year 2008. For the stochastic SFA and StoNED methods, the CE is based on the formulation of the conditional expected value of the inefficiency term u, developed by Jondrow et al. (1982) (henceforth referred to as the JLMS method). The total cost reduction target for the whole industry is reported in the first column of Table 8 (all figures in €1000 at prices of the year 2008). The remaining columns provide summary statistics of the firm-specific cost reduction targets.

Comparison of the cost reduction targets reveals substantial differences between the methods considered. The calculated total cost reduction target of the industry based on the StoNED estimates is somewhat lower than €50 million. The SFA estimate is approximately €50 million larger than the corresponding StoNED figure. Further, the DEA estimate is approximately €50 million larger than the SFA estimate. Although the efficiency scores obtained with the different methods are highly correlated, the monetary figures presented in Table 8 illustrate that the choice of the estimation method does have a significant economic impact within the regulatory framework.

Based on Table 8, one might conclude that StoNED is most favorable to the regulated firms, whereas DEA is the best method from the perspective of consumers. However, the practical implementation of the cost reduction targets in the regulatory framework will also matter. In the previous regulation period, EMV made several adjustments to the cost reduction targets to make them more favorable to the regulated industry. In the current regulation period EMV enforces the StoNED targets more vigorously. While a detailed discussion of the practical implementation of the EMV model falls beyond the scope of the present paper, in the next section we do discuss two important insights gained during the reform of the EMV model.

5. Implementation of efficiency benchmarks in regulation

Estimation of firm specific efficiency scores is often the main objective of frontier estimation. Consequently, the benchmark regulation typically starts from the efficiency scores. In Finland, EMV is required to provide firm specific efficiency scores as the basis of regulation by the law. In this section we argue that the frontier cost function provides more appropriate benchmarks for the acceptable cost level or cost reduction targets.

In the deterministic models such as DEA, the production technology can be fully characterized by the distance to frontier (i.e., the distance function). In this case, the efficiency scores can be harmlessly used for setting cost reduction targets. However, the situation is different in the stochastic models such as SFA and StoNED because the distance to frontier is subject to random noise. Even though SFA is currently used in regulation in some countries, the impact of noise has not been recognized. Two important lessons from the recent reform of the Finnish regulatory model by EMV are worth noting.

First, we emphasize that the estimation of the stochastic cost frontier function rests on a much sounder statistical foundation than the estimation of firm specific efficiency scores. Provided that the model assumptions hold, the cost frontier can be consistently estimated even in a cross-sectional setting subject to noise. In contrast, it is well known in SFA literature that the firm specific inefficiency estimates obtained by using the JLMS method are inconsistent. The rationale of this argument can be stated as follows. The frontier cost function is common to all firms, and hence the noise contained in individual observations can be averaged out. In contrast, firm specific efficiency estimates are based on the distance from an individual observation to the frontier. Even if the sample size increases, the distance is measured from a single data point to the frontier. The increase in sample size generally improves the precision of the frontier estimator, but the efficiency estimator is still based on the distance of a single data point to the frontier, and hence the noise contained in the single data point cannot be averaged out.[10]

Second, the cost reduction targets based on the JLMS method are dynamically inconsistent, as first noted by Kuosmanen et al. (2010). The argument can be briefly stated as follows. The JLMS method transforms the distance to the frontier to conditional expected value of inefficiency. As a result, it attributes some proportion of the measured distance to the frontier to the noise term. The larger the distance to frontier, the larger the assumed impact of noise. In another context, Wang and Schmidt (2002) refer to this as the shrinkage effect of the JLMS method. In the present context, the dynamic inconsistency arises from the fact that the regulated firm does not necessarily reach the frontier even if it improves its efficiency by the amount suggested by the JLMS method. Even if all firms in the regulated industry improve their efficiency according to the JLMS method, there is no guarantee

[10] In the case of panel data where n firms are observed over T time periods, it is possible to estimate time invariant inefficiency by averaging out noise over the T observations of the same firm. To estimate the cost frontier we can average out noise over the full sample of nT observations, which will likely result as a more precise estimator. Further, the consistent estimation of the inefficiency term in the panel data setting requires some additional assumptions, which may be considered restrictive. For example, one could assume a time invariant inefficiency term or a specific functional form for the efficiency change over time. For a freely time varying inefficiency estimator that does not impose any additional assumptions, consistent estimation is not possible even in the panel data setting.

that all the firms reach the frontier at the end of the regulation period.

To address both problems noted above, in the current regulation period EMV defines cost efficiency as the ratio of the frontier cost function and the observed cost of the firm. Specifically,

$$CE' = C(y_1, y_2, y_3) \times \exp(\delta z)/x \quad (3)$$

In this measure, the nominator can be consistently estimated, avoiding the inconsistency of the JLMS estimator. Note that the denominator x contains both inefficiency and noise. However, the presence of noise in the denominator is not a problem as our objective is to specify efficiency improvement targets such that the firms would reach the efficient cost level $C(y_1, y_2, y_3) \times \exp(\delta z)$. Indeed, the acceptable cost level defined using Eq. (3) is dynamically consistent: if a firm reduces its current cost level by factor *CE'* during the regulation period, it will reach the efficient cost frontier at the end of the regulation period.

6. Monte Carlo simulations

The empirical comparison presented in the previous section shows that the choice of the frontier estimation method does matter in the regulation. We next examine performance of alternative methods in a simulated setting where the true cost frontier and the firm-specific inefficiencies are known beforehand. The advantage of the Monte Carlo (MC) comparison is that it allows us to quantify the performance of each method in terms of standard criteria such as the bias and root mean squared error (to be defined below).

A critical step in the MC analysis is the specification of the data generating process (DGP) that produces the simulated data. For the empirical relevance of the MC analysis, it is desirable to specify the DGP to imitate both the characteristics of the regulatory model the observed patterns of empirical data. In this study, we calibrate the DGP to reflect both these aspects.

6.1. Data generating process (DGP)

The generic cost frontier Model (1) forms the basis of our DGP. To ensure comparability, in the MC comparisons we apply exactly the same model specification across all methods. Thus, we assume a three output case and omit the contextual variable z, as its proper inclusion in DEA would be somewhat more complicated than in SFA or StoNED.

We first generate random data for the three output variables using the formulas presented in Table 9. The DGP for output variables has been specified to mimic the observed data as closely as possible. The empirical distribution of the logarithms of outputs is approximately uniformly distributed within the range [3, 11]. First, we generate the data for transmitted energy. The other two outputs are generated conditional on the first output. A weighted average of a random draw from uniform distribution and the previously generated energy output is applied to generate these variables such that the weights are based on the empirical correlation between the observed variables. For example, the empirical correlation between the network length and the transmitted energy is 0.87. Thus, the simulated output data exhibit similar correlations as the observed output variables in our empirical data.

Table 9
DGP for the output variables.

Output	DGP
Energy	$y_{1,i} = \exp(\mathrm{Uni}[3, 11])$
Network length	$y_{2,i} = \sqrt{(1-0.87^2)} \times \exp(\mathrm{Uni}[3, 11]) + 0.87 \times y_{1,i}$
Customers	$y_{3,i} = \sqrt{(1-0.98^2)} \times \exp(\mathrm{Uni}[3, 11]) + 0.98 \times y_{1,i}$

Given the simulated output data, the next step is to generate the total cost. This requires a specification of the cost function. Recall that the commonly used functional forms such as the Cobb–Douglas and translog are inappropriate in the present context. To calibrate our DGP to the current regulatory practice of EMV as closely as possible, we apply the piece-wise linear cost frontier applied by EMV in the regulation period 2012–2015. Given the output vector $(y_{1,i}, y_{2,i}, y_{3,i})$, the value of the cost frontier is calculated as

$$C_i = \max_h(\beta_{1h} y_{1i} + \beta_{2h} y_{2i} + \beta_{3h} y_{3i}) \quad (4)$$

where $(\beta_{1h}, \beta_{2h}, \beta_{3h})$, $h = 1, \ldots, H$ are the slope coefficients (marginal costs) of the H different hyperplane segments of the piece-wise linear cost frontier implemented by EMV (compare with the shadow prices reported in Table 2 and problem (2) in the technical appendix provided in the Supplement). Note that the max operator in Eq. (4) selects the most favorable output prices for each simulated data point.

Having calculated the values of the frontier cost function (which represents the efficient cost level) for each simulated point, the observed total cost are generated using

$$x_i = C_i \times \exp(u_i + v_i), \quad (5)$$

where the inefficiency u for the noise v are distributed as: $u_i \sim |N(0, 0.17^2)|$ and $v_i \sim N(0, 0.09^2)$. The parameter values of the standard deviations of the inefficiency and noise terms are calibrated based on the empirical estimates obtained by applying the method of moments estimator to the CNLS residuals in the StoNED procedure.

Before proceeding to the results, it is worth to discuss whether and to what extent the DGP provides an unfair advantage to any of the methods considered. First, the DGP does not violate any of the assumptions of the StoNED method. The piece-wise linear functional form of the true cost function used in the simulations is compatible with the form of the StoNED frontier, but the same is true for DEA. The fact that the coefficients $(\beta_{1h}, \beta_{2h}, \beta_{3h})$ and the parameters (σ_u, σ_v) have been *ex ante* estimated by the StoNED method does not give any particular advantage to this or that method: the purpose of the *ex ante* estimation is to match the DGP with the current regulatory practice of EMV. As for DEA, the presence of the noise term v violates the deterministic nature of this method. However, empirical data are always subject to some noise, and some authors explicitly suggest that DEA is robust enough to tolerate some noise (Gstach, 1998; Banker and Natarajan, 2008). In fact, the noise term can help to alleviate the small sample bias of the DEA estimator, as we note below. Regarding SFA, the piece-wise linear functional form violates the maintained assumption of the linear cost function. In all other respects, the SFA estimator is correctly specified: we assume the half-normal distribution of the inefficiency term (in contrast to the EMV specification of truncated normal inefficiency used in the previous sections). For comparability of SFA and StoNED, we apply the MOLS estimation strategy for SFA and the method of moments estimator in StoNED.

The DGP used in the present simulations may seem to favor StoNED, as it is the only method with the assumptions consistent with those of the DGP. However, the rigid functional form of SFA and the deterministic orientation of DEA are the well-known characteristics of these methods. We must also stress that the NMA approach is supposed to remedy these issues. Hence, we find it meaningful to compare the performances of the methods using the DGP described above. For further Monte Carlo comparisons of DEA, SFA and StoNED under alternative data generation processes (including smooth frontiers and scenarios without noise), a reader

is referred to Kuosmanen and Kortelainen (2012), Johnson and Kuosmanen (2012), and Andor and Hesse (in press).

6.2. Performance measures

Recall from Section 3 that the SFA and StoNED estimators of the cost frontier C are consistent, whereas the JLMS estimator of firm-specific inefficiency is inconsistent. Since no consistent estimator of firm-specific inefficiency is available in the stochastic setting involving noise, we compare performance of the methods in terms of their precision in estimating the cost frontier C. Given the simulated values C_i (calculated using Eq. (4)) and the corresponding estimates $\hat{C}_i$ (obtained with StoNED, DEA, SFA, and NMA), the performance of the method is measured using the root mean squared error (RMSE) and bias, defined as

$$\text{RMSE} = \frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{C}_i - C_i}{C_i}\right)^2} \tag{6}$$

$$\text{BIAS} = \frac{1}{M}\frac{1}{n}\sum_{m=1}^{M}\sum_{i=1}^{n}\frac{\hat{C}_i - C_i}{C_i} \tag{7}$$

where M denotes the number of replications in the simulation. Note that the RMSE is always greater than or equal to zero, with zero indicating perfect precision. In contrast, the bias can be positive or negative, the positive values indicating overestimation and the negative values underestimation of the cost function. For both performance statistics, values close to zero are desirable. Both RMSE and bias have been normalized such that the performance statistics have an interpretation as an average dispersion or bias. For example, RMSE=0.05 indicates that the estimates $\hat{C}_i$ deviate from the true C_i value by 5% on average.

6.3. Simulation results

The MC simulations were conducted using the GAMS software and the MINOS solver run on a standard desktop PC (the GAMS code for the simulations is available as online annex). We consider four different scenarios with sample sizes $n=25$, 50, 100, and 200. The sample sizes are chosen to be relatively small to reflect the usual number of firms in this kind of sector (in the EMV data, $n=89$). Each scenario has been replicated $M=1000$ times. The results of the MC analysis are reported in Table 10.

Consider first the RMSE statistics reported on the left panel of Table 10. The StoNED estimator has a lower RMSE than other methods at all sample sizes. The average dispersion of approximately 5% from the true value is a very good result in the stochastic setting involving noise. Note that the precision of the StoNED estimator improves (RMSE decreases) as the sample size increases, as expected. The DEA estimator yields a relatively good precision of RMSE less than 10% at small sample sizes. However, the RMSE increases together with the sample size. This is due to the fact that DEA ignores the noise term. In small samples, the noise term and the small sample bias offset each other, but as the sample size increases, the bias due to the noise term starts to dominate. The SFA estimator yields catastrophic results in this comparison, with average deviations of the magnitude of 50–200%. Recall that the linear functional form is severely wrongly specified in these simulations; most reported MC simulations assume the correct (or almost correct) functional form for SFA. It is not surprising to find that the linear functional form fails to capture the piece-wise linear cost function, Eq. (4), used in our simulations. Further, the high correlation between the output variables makes SFA vulnerable to multicollinearity. Moreover, note that the RMSE of SFA increases alarmingly as the sample size increases. Finally, the MC simulations illustrate the weakness of the NMA approach: the poor performance of SFA carries over to the NMA estimator. In this case, the use of DEA alone is clearly superior to NMA.

Table 10
Simulation results.

	RMSE				BIAS			
	$n=25$	$n=50$	$n=100$	$n=200$	$n=25$	$n=50$	$n=100$	$n=200$
StoNED	0.072	0.057	0.044	0.027	0.030	0.022	0.014	0.009
DEA	0.088	0.091	0.107	0.129	−0.025	−0.060	−0.091	−0.118
SFA	0.469	0.886	1.439	1.923	−0.253	−0.666	−1.192	−1.661
NMA	0.254	0.464	0.750	1.003	−0.139	−0.363	−0.641	−0.890

The bias statistics are reported on the right panel of Table 10. The bias of the StoNED estimator is small, and decreases as the sample size increases. In contrast to DEA and SFA, the bias of the StoNED estimator is positive, which means that StoNED tends to overestimate the true cost level in this setting. In the context of regulation, modest overestimation is generally preferred to underestimation. The conventional wisdom of DEA suggests that the DEA estimator is systematically biased towards overestimation of cost. However, this idea stems from the deterministic setting, whereas in the present MC simulations the DGP contains noise. The results of Table 10 aptly illustrate that the DEA estimator is downward biased under noise. In very small samples, the noise term can offset the small sample bias, as we noted above.

Finally, we must emphasize that the previous MC comparison has been calibrated to mimic the regulatory model of EMV and the empirical data of the Finnish electricity distribution firms as closely as possible. The purpose of such tailored simulations is to ensure the relevance of the MC evidence in the specific context of the Finnish regulatory model. We stress that the results of this section do not necessarily apply to other sectors or in other countries. As MC simulations are nowadays relatively inexpensive, we suggest that investigating the internal consistency the benchmarking methods through MC simulations calibrated to the specific regulatory context should be routinely conducted.

7. Conclusions

In this paper we have compared the frontier estimation techniques applied in the benchmark regulation of electricity distribution firms. The comparison was conducted both in terms of the empirical data from Finland and in the controlled environment of Monte Carlo simulations. Our empirical comparison demonstrated that the choice of benchmarking method has significant economic effects on the regulatory outcomes, even when the efficiency estimates from different methods are highly correlated. Although the frontier estimation methods are often used for assessing relative efficiency and ranking of firms, in the context of regulation, also the level of efficiency matters.

A unique feature of our Monte Carlo simulations concerns the specification of the data generating process. We calibrated the simulation model and its parameters to capture as closely as possible the key characteristics of the distribution sector and the regulatory system in Finland. This allows us to estimate the potential bias and dispersion of the frontier estimates obtained with different frontier estimation methods in the setting that mimics the empirical reality of this sector.

We have learned at least five important lessons from this study:

(1) *Heterogeneity:* a large proportion of the observed dispersion in cost per kilowatt hours across firms can be explained and attributed to the heterogeneity of firms and their operating

environments. The benchmarking model should be flexible enough to take into account the different circumstances of small firms and large corporations, firms operating in rural area or in a large city, and firms that supply power to households or heavy industry. The current regulatory model of EMV attempts to take the heterogeneity into account through the application of a non-parametric piece-wise linear cost frontier that allows the marginal costs of outputs differ across firms, and by applying the proportion of underground cables as an additional contextual variable.

(2) *Noise:* the cost data are subject to random variation from various sources. For example, the capital expenditures depend on somewhat arbitrary accounting rules and depreciation rates. Random weather events such as storms cause interruptions, which influence the operational costs. In these circumstances, stochastic frontier models that explicitly recognize a random noise term are preferable to deterministic benchmarks that attribute all deviations from the frontier to inefficiency. The current regulatory model of EMV takes a random noise term explicitly and systematically into account in the frontier estimation.

(3) *Frontier as the benchmark:* it is important to recognize that the estimation of the frontier cost function (or production function) rests on a much sounder statistical foundation than the estimation of firm-specific efficiency scores. Therefore, it is generally recommended to set the efficiency improvement targets based on the frontier cost or production function, rather than the firm-specific efficiency estimates, as noted in Section 5. In the current regulatory model of EMV, efficiency is defined as the ratio of the cost frontier and the observed cost, which effectively imposes the cost frontier as the target.

(4) *Implementation:* development of a benchmarking model should not be viewed as an isolated exercise, but rather as an integral part of designing the regulatory framework as a whole. The systematic use of the StoNED cost frontier as a benchmark has enabled EMV to eliminate some redundant components in the regulatory model, making the regulation more transparent. Although the efficiency estimates according to the StoNED method are on average higher than those of DEA and SFA, EMV has implemented the efficiency improvement targets more vigorously than in the previous regulation periods.

(5) *Tailored simulations:* in this paper we have shown that it is possible to calibrate the simulation model to mimic the characteristics of the regulated industry as well as the regulatory model. Conducting tailored simulations is an inexpensive way to compare the performance of alternative benchmarking tools in the specific context of application. We would recommend the use of calibrated Monte Carlo simulations as a test for the internal consistency of the chosen benchmarking model.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.enpol.2013.05.091.

References

Agrell, P., Bogetoft, P., 2007. Development of benchmarking models for German electricity and gas distribution, Sumicsid AB, Project GERNER/AS6, Final Report. ⟨http://www.bundesnetzagentur.de/cae/servlet/contentblob/88060/publicationFile/1932/GutachtnSUMICSID-Id9598.pdf⟩ (accessed 12.10.11).

Agrell, P., Bogetoft, P., Tind, J., 2005. DEA and dynamic yardstick competition in Scandinavian electricity distribution. Journal of Productivity Analysis 23, 173–201.

Aigner, D.J., Lovell, C.A.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier models. Journal of Econometrics 6, 21–37.

Andor, M., Hesse, F. The StoNED age: the departure into a new era of efficiency analysis? A Monte Carlo comparison of StoNED and the "oldies" (SFA and DEA). Journal of Productivity Analysis, in press.

Azadeh, A., Ghaderia, S.F., Omranib, H., Eivazya, H., 2009. An integrated DEA–COLS–SFA algorithm for optimization and policy making of electricity distribution units. Energy Policy 37 (7), 2605–2618.

Banker, R.D., Natarajan, R., 2008. Evaluating contextual variables affecting productivity using data envelopment analysis. Operations Research 56 (1), 48–58.

Bogetoft, P., Otto, L., 2011. Benchmarking with DEA, SFA, and R. International Series in Operations Research and Management Science. 157. Springerhttp://dx.doi.org/10.1007/978-1-4419–7961-2_10.

Charnes, A., Cooper, W.W., Rhodes, E., 1978. Measuring the efficiency of decision making units. European Journal of Operational Research 2, 429–444.

Cullmann, A., 2009. Parametric and nonparametric efficiency analysis in electricity distribution: a European perspective. Ph.D. Dissertation, Technische Universitat Berlin. ⟨http://opus.kobv.de/tuberlin/volltexte/2009/2192/pdf/cullmann_astrid.pdf⟩ (accessed 24.05.12).

Cullmann, A., 2012. Benchmarking and firm heterogeneity: a latent class analysis for German electricity distribution companies. Empirical Economics 42 (1), 147–169.

Dassler, T., Parker, D., Saal, D.S., 2006. Methods and trends of performance benchmarking in UK utility regulation. Utilies Policy 14, 166–174.

Eskelinen, J., Kuosmanen, T. Intertemporal efficiency analysis of sales teams of a bank: stochastic semi-nonparametric approach. Journal of Banking and Finance, http://dx.doi.org/10.1016/j.jbankfin.2013.03.010, in press.

Farrell, M.J., 1957. The measurement of productive efficiency. Journal of the Royal Statistical Society A 120, 253–281.

Forsund, F.R., Kittelsen, S.A.C., 1998. Productivity development of Norwegian electricity distribution utilities. Resources and Economics 20, 207–224.

Fried, H.O., Lovell, C.A.K., Schmidt, S.S. (Eds.), 2008. Measurement of Productive Efficiency and Productivity Growth, second ed. Oxford University Press, New York.

Greene, W.H., 2005. Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. Journal of Econometrics 126, 269–303.

Gstach, D., 1998. Another approach to data envelopment analysis in noisy environments: DEA+. Journal of Productivity Analysis 9, 161–176.

Hjalmarsson, L., Veiderpass, A., 1992. Productivity in Swedish electricity retail distribution. Scandinavian Journal of Economics 94, 193–205.

Iglesias, G., Castellanos, P., Seijas, A., 2010. Measurement of productive efficiency with frontier methods: a case study for wind farms. Energy Economics 32, 1199–1208.

Irastorza, V., 2003. Benchmarking for distribution utilies: a problematic approach to defining efficiency. Electricity Journal 16 (10), 30–38.

Jamasb, T., Nillesen, P., Pollitt, M., 2003. Gaming the regulator: a survey. Electricity Journal 16 (10), 68–80.

Jamasb, T., Nillesen, P., Pollitt, M., 2004. Strategic behaviour under regulatory benchmarking. Energy Economics 26, 825–843.

Jamasb, T., Pollit, M., 2001. Benchmarking and regulation: international electricity experience. Utilies Policy 9, 107–130.

Jamasb, T., Pollit, M., 2003. International benchmarking and regulation: an application to European electricity distribution utilizes. Energy Policy 31, 1609–1622.

Jondrow, J., Lovell, C.A.K., Materov, I.S., Schmidt, P., 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. Journal of Econometrics 19 (2-3), 233–238.

Johnson, A.L., Kuosmanen, T., 2011. One-stage estimation of the effects of operational conditions and practices on productive performance: asymptotically normal and efficient, root-n consistent StoNED method. Journal of Productivity Analysis 36, 219–230.

Johnson, A.L., Kuosmanen, T., 2012. One-stage and two-stage DEA estimation of the effects of contextual variables. European Journal of Operational Research 220 (2), 559–570.

Kinnunen, K., 2006. Investment incentives: regulation of the Finnish electricity distribution. Energy Policy 34, 853–862.

Koopmans, T.C., 1951. An analysis of production as an efficient combination of activities. In: Koopmans, T.C. (Ed.), Activity Analysis of Production and Allocation. Cowles Commission for Research Monograph, New York, p. 13.

Kopsakangas-Savolainen, M., Svento, R., 2008. Estimation of cost-effectiveness of the Finnish electricity distribution utilities. Energy Economics 30 (2), 212–229.

Korhonen, P., Syrjänen, M., 2003. Evaluation of cost efficiency in Finnish electricity distribution. Annals of Operations Research 121, 105–122.

Kumbhakar, S.C., Lovell, C.A.K., 2000. Stochastic Frontier Analysis. Cambridge University Press, New York.

Kuosmanen, T., 2012. Stochastic semi-nonparametric frontier estimation of electricity distribution. Energy Economics 34 (6), 2189–2199.

Kuosmanen, T., Johnson, A.L., 2010. Data envelopment analysis as nonparametric least squares regression. Operations Research 58 (1), 149–160.

Kuosmanen, T., Kortelainen, M., Kultti, K., Pursiainen, H., Saastamoinen, A., Sipiläinen, T., 2010. Sähköverkkotoiminnan kustannustehokkuuden estimointi StoNED-menetelmällä: ehdotus tehostamistavoitteiden ja kohtuullisten kustannusten arviointiperusteiden kehittämiseksi kolmannella valvontajaksolla 2012–2015 (in Finnish). Available from: ⟨www.emvi.fi⟩ (accessed 13.12.11).

Kuosmanen, T., Kortelainen, M., 2012. Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. Journal of Productivity Analysis 38 (1), 11–28.

Kuosmanen, T., Kuosmanen, N., 2009. Role of benchmark technology in sustainable value analysis: an application to Finnish dairy farms. Agricultural and Food Science 18 (3–4), 302–316.
Laffont, J.-J., Tirole, J., 1993. A Theory of Incentives in Procurement and Regulation. MIT Press, Cambridge, MA.
Mekaroonreung, M., Johnson, A.L., 2012. Estimating the shadow prices of SO_2 and NO_x for U.S. coal power plants: a convex nonparametric least squares approach. Energy Economics 34 (3), 723–732.
Olson, J.A., Schmidt, P., Waldman, D.M., 1980. A Monte Carlo study of estimators of stochastic frontier production functions. Journal of Econometrics 13, 67–82.
Pollit, M., 2005. The role of efficiency estimates in regulatory price reviews: Ofgem's approach to benchmarking electricity networks. Utilies Policy 13, 279–288.
Shleifer, A., 1985. A theory of yardstick competition. Rand Journal of Economics 16 (3), 319–327.
Simar, L., Wilson, P.W., 2008. Statistical inference in nonparametric frontier models: recent developments and perspectives. In: Fried, H.O., Lovell, C.A.K., Schmidt, S. S. (Eds.), Measurement of Productive Efficiency and Productivity Growth, 2nd edition Oxford University Press.
Syrjänen, M., Bogetoft, P., Agrell, P., 2006. Analogous efficiency measurement model based on stochastic frontier analysis. Gaia Consulting Oy. Available from: ⟨www.energiamarkkinavirasto.fi⟩ (accessed 11.12.06).
Tsionas, M., 2002. Stochastic frontier models with random coefficients. Journal of Applied Econometrics 17, 127–147.
Wang, H., Schmidt, P., 2002. One step and two step estimation of the effects of exogenous variables on technical efficiency levels. Journal of Productivity Analysis 18, 129–144.
Weyman-Jones, T.G., 1991. Productive efficiency in a regulated industry: the area electricity boards of England and Wales. Energy Economics 13 (2), 116–122.

Supplement to the article:

What is the Best Practice for Benchmark Regulation of Electricity Distribution? Comparison of DEA, SFA and StoNED Methods:

Timo Kuosmanen[1,*], Antti Saastamoinen[1], Timo Sipiläinen[2]

1) School of Business, Aalto University, 00076 Aalto, Helsinki, Finland. E-mail. Firstname.Lastname@aalto.fi.
2) Department of Economics and Management, University of Helsinki, P.O. Box 27, 00014 Helsinki, Finland. E-mail. Firstname.Lastname@helsinki.fi.
*) Corresponding author: Tel. +358 9 43131, Fax. +358 943 138700.

Technical Appendix

This technical supplement introduces the basic characteristics and discusses the properties of the three frontier estimators examined in the abovementioned paper. The purpose of this technical supplement is to preserve space for the comparison of estimated empirical frontiers instead of their theoretical properties. For simplicity of the presentation, we first state the general cost frontier model and then discuss each model in turn. We also deal with the naïve model averaging (NMA) approach here in more detail and discuss the problems of this approach on a technical level.

The general cost frontier model

Here we briefly state the general cost frontier model. The cost regulation of EMV is based on the following generic model of cost frontier (see Kuosmanen, 2012, for a more detailed discussion).

$$\ln x = \ln C(y_1, y_2, y_3) + \delta z + u + v \quad (1)$$

where

x is the observed total cost (TOTEX) (1,000 €)
C is the frontier cost function
y_1 is the energy transmission (GWh)
y_2 is the total length of the network (km)
y_3 is the number of customers
z is the proportion of underground cables
δ is the coefficient of the z variable
u is the random variable representing inefficiency
v is the random variable representing stochastic noise

In this study the cost variable x refers to the total expenditure (TOTEX), which consists of three components: controllable operational costs (OPEX), capital expenditures (CAPEX) and the external supply interruption costs for customers (INT). The last component can also be viewed as a quality component, as the lack of supply interventions can be interpreted as an indicator of good service.

StoNED estimator

The StoNED estimator is based on some general axioms (or regularity conditions) concerning the benchmark technology (see details of the method in Kuosmanen and Kortelainen, 2012). The set of axioms imposed in the EMV regulatory model are the following (see Kuosmanen, 2012):

1) C is monotonic increasing in all outputs
2) C is globally convex in outputs
3) C exhibits constant returns to scale (CRS)

The first two conditions are standard properties in DEA. The third axiom defines the nature of returns to scale and it could be relaxed. However, the CRS axiom could not be rejected in the empirical specification test reported by Kuosmanen (2012). More importantly, the CRS axiom is preferable from the regulatory point of view, as the benchmark technology exhibits the same level of total factor productivity irrespective of the firm size. For example, suppose firms enjoy economies of scale in reality. The CRS axiom of the regulatory model then provides an incentive for firms to seek productivity improvement through mergers. Such an incentive would be lost if the CRS axiom were relaxed and variable returns to scale (VRS) were imposed. Indeed, the use of the VRS benchmark may give wrong incentives for firms to split or merge for strategic reasons to game the regulator (see e.g. Jamasb et al., 2003, 2004).

The StoNED estimation proceeds in two stages. First, the cost frontier model (1) is estimated with convex nonparametric least squares (CNLS: Johnson and Kuosmanen, 2011, 2012; Kuosmanen, 2008). Denoting the composite error term by $\varepsilon_i = u_i + v_i$, the CNLS problem can be stated as

$$\min_{\phi,\beta,\delta,\varepsilon} \sum_{i=1}^{n} \varepsilon_i^2$$

subject to (2)

$$\begin{aligned}
&\ln x_i = \ln \phi_i + \delta z_i + \varepsilon_i \quad \forall i \\
&\phi_i = \beta_{1i} y_{1i} + \beta_{2i} y_{2i} + \beta_{3i} y_{3i} \quad \forall i \\
&\phi_i \geq \beta_{1h} y_{1i} + \beta_{2h} y_{2i} + \beta_{3h} y_{3i} \quad \forall h,i \\
&\beta_{ki} \geq 0 \ \forall k = 1,2,3; \forall i
\end{aligned}$$

The firm specific beta coefficients represent the marginal costs of outputs (shadow prices). Alternatively, these coefficients can be interpreted as the slopes of the tangent hyperplanes to the piece-wise linear cost frontier. The firm specific coefficients allow for greater heterogeneity of distribution networks than the usual parametric approaches (cf., e.g., Cullmann, 2012). The contextual variable z also partly captures heterogeneity of firms.

In the second stage we impose distributional assumptions on inefficiency and noise and follow the method of moments approach (see Kuosmanen and Kortelainen, 2012; Kuosmanen, 2012) to estimate the firm specific inefficiencies, utilizing the estimated CNLS residuals $\hat{\varepsilon}_i$. The usual assumptions of the SFA literature are that u_i has a half-normal distribution such that $u_i \geq 0$, and v_i has a normal distribution with zero-mean and a finite constant variance, and these assumptions are assumed here also.

The optimal ϕ_i from problem (2) is a consistent estimator of the total cost x_i, conditional on outputs (y_{1i}, y_{2i}, y_{3i}), that is,

$$E(x_i \mid y_{1i}, y_{2i}, y_{3i}) = C(y_{1i}, y_{2i}, y_{3i}) \times \exp(\mu), \qquad (3)$$

where μ is the expected inefficiency estimated with methods of moments, given the distributional assumption of inefficiency. This approach is similar to the Modified OLS (MOLS) approach commonly applied in SFA literature (Aigner et al., 1977; Olson et al., 1980).[1] Indeed, StoNED can be seen as an axiomatic, nonparametric variant of the classic MOLS; the conditional expected value $E(x_i \mid y_{1i}, y_{2i}, y_{3i})$ is estimated by CNLS instead of OLS, but otherwise the StoNED estimator follows the standard MOLS procedure.

To estimate the frontier cost function, we must adjust the estimated ϕ_i with μ. Thus, the StoNED cost frontier is obtained by adjusting the estimated ϕ_i downward according to

$$\hat{C}^{StoNED}(y_{1i}, y_{2i}, y_{3i}) = \phi_i \times \exp(-\hat{\sigma}_u \sqrt{2/\pi}). \qquad (4)$$

Finally, we can utilize the Jondrow et al. (1982) decomposition to obtain firm-specific inefficiency estimates $\hat{u}_i$. For comparability with the DEA efficiency scores, we convert the inefficiency estimates as cost efficiency measures as follows

$$CE_i = 100\% \times \exp(-\hat{u}_i) \qquad (5)$$

[1] Sometimes MOLS is referred to as corrected OLS (COLS) (see, e.g., Azadeh et al., 2009). We prefer to use MOLS for the probabilistic estimator that takes into account noise, and reserve the term COLS for the deterministic estimator that envelopes all observations.

In practice, the CNLS problem (2) can be solved by mathematical programming solvers for convex problems. In this study we use GAMS (*General Algebraic Modeling System*) and its MINOS solver as this solver is suitable for solving nonlinear programming problems. Problem (2) is nonlinear due the logarithmic transformations applied to the observed costs and the estimable frontier costs. Since there is a large number of constraints and parameters, problem (2) is computationally more burdensome than for example the OLS. With the present hardware and software capacity, however, problem (2) is solvable in tolerable time by standard PC, provided that the sample size is not too large (see Lee et al., 2013, for discussion).

DEA estimator

The DEA estimator can be obtained as restricted special case of Problem (2). If we restrict the residuals $\hat{\varepsilon}_i$ to take only positive values and exclude the contextual variable z, the CNLS problem (2) is equivalent to the input-oriented DEA under CRS (see Kuosmanen and Johnson, 2010, for details). Thus, DEA maintains the same assumptions concerning the shape of the frontier as StoNED.

If we assume away noise, the DEA estimator is consistent, but biased in the small samples (Banker, 1993). In the case of the cost frontier, DEA overestimates the true unobserved cost function in the small samples but it converges to the true frontier as the sample size tends towards infinity. Statistical inference on DEA can be conducted by using the bootstrap methods (e.g., Simar and Wilson, 2008). However, if the stochastic noise term is included in the model, the DEA estimator can be biased in both directions. In this case the bootstrap inferences are invalid. Indeed, it seems a common misunderstanding to assume that the bootstrap method (or robust frontiers) would make DEA more robust to noise. We must emphasize that the probabilistic treatment of sampling error does not address stochastic noise at all.

The EMV specification of the DEA model applied in the previous regulation period 2008 – 2011 did not include any contextual variables z. The conventional approach to modeling z-variables in DEA is

to resort to a two-stage approach, where efficiency is first estimated using DEA, and then the DEA efficiency scores are regressed on z-variables, using OLS, probit, tobit, or truncated regression. Simar and Wilson (2007) present heavy critique of this approach. Recently, Johnson and Kuosmanen (2012) have shown that one-stage estimation of z-variables is possible in DEA. However, we follow the EMV specification and omit the z-variable from DEA altogether.

SFA estimator

Within this context, the SFA estimator of the frontier can be obtained as special case of StoNED estimator if cost frontier C is assumed to be linear (as in Syrjänen et al., 2006, specification implemented by EMV in 2008 – 2011) and we restrict the marginal costs to be same for every firm (i.e., $\beta_{ki} = \beta_{kh}\ \forall i,h,k$). The estimation of inefficiency in SFA is analogous to the procedure presented above for StoNED as StoNED lends its approach from SFA.

The Finnish Energy Market Authority attempted to take the heterogeneity of firms into account in SFA by dividing the network variable into two separate variables. The use of different sets of output variables in the DEA and SFA models is however problematic for the parallel use of both methods as a part of the regulatory model. This issue is discussed in next section of this appendix when we deal the NMA approach. It is also good to note that this is not the only way to take account heterogeneity in SFA. The SFA literature offers abundant number of ways for modeling contextual variables z (e.g., Kumbhakar and Lovell, 2000, Ch. 7, and references therein). However, again we restrict ourselves to the EMV specification with divided network variable.

Naïve model averaging (NMA)

Given the relative strengths and limitations of DEA and SFA, it might be tempting to try to alleviate the risk of model misspecification by taking the average of the two estimators. In Finland, EMV applied the average of DEA and SFA estimators in the previous regulation period 2008 – 2011. Consequently, we refer

this simplistic approach as naïve model averaging (NMA). This section provides a brief but critical examination of the shortcomings of NMA.

Let us first consider the statistical properties of NMA based on the known properties of SFA and DEA. If the parametric assumptions of the SFA estimator hold, both the MOLS and the maximum likelihood estimators of the cost frontier C are unbiased and consistent (Greene, 2008). The firm specific inefficiency term u_i can be estimated by using the conditional expected value of Jondrow et al. (1982). This estimator is unbiased, but inconsistent. In the cross-sectional setting, the inconsistency of the firm-specific inefficiency estimator is due to the fact that inefficiency is estimated based on the residuals and there is only one observation available for each firm. While an increase in the sample size improves the fit of the cost frontier, it does not improve precision of the firm-specific efficiency estimates. Thus, if we are interested in firm-specific efficiency scores, then inconsistency of the SFA estimator directly implies the NMA estimator is inconsistent even if the assumptions of the SFA model hold.

To obtain a consistent estimator of firm-specific efficiency, we must assume away noise. In this case, the DEA estimator is consistent under the stated axioms. The SFA estimator remains inconsistent even if the functional form is correctly specified, so there is little benefit to introduce SFA: the DEA estimator is consistent, whereas NMA is not. But by assuming away the noise, we lose the most desirable property of SFA.

As for the estimation of the cost frontier C, the statistical consistency of the NMA estimator requires that the assumptions of both DEA and SFA hold simultaneously. That is, the NMA estimator is consistent only if the frontier is linear with respect to outputs, inefficiency u has a truncated normal distribution, and there is no noise v. In this situation SFA estimator is unbiased and consistent. The DEA estimator is consistent but biased. Thus, the NMA estimator is consistent but biased. We conclude that under the assumptions required for the statistical consistency of the NMA cost frontier estimator, the SFA

estimator is both unbiased and more efficient than the NMA estimator: introducing the DEA estimator does not provide any real benefit in this situation.

The problems of NMA are further intensified by the fact that EMV applied different sets of output variables in DEA and SFA. In DEA the total network length was used as an output, whereas in SFA the network length was divided in two output variables, the urban underground cabled network and other network. This creates a profound misspecification problem. If the two models are differently specified with respect to the output variables, then one of the models (if not both) has to be misspecified. If one of the models is misspecified, then so is the NMA estimator. There is no reason to expect that averaging wrongly specified estimators would be beneficial.

REFERENCES

Aigner, D.J., Lovell, C.A.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier models. *Journal of Econometrics*, 6, 21-37.

Azadeh, A., Ghaderia, S.F., Omranib, H., Eivazya, H. 2009 An integrated DEA–COLS–SFA algorithm for optimization and policy making of electricity distribution units. *Energy Policy*, 37 (7), 2605-2618.

Banker, R.D., Natarajan, R., 2008. Evaluating contextual variables affecting productivity using data envelopment analysis. *Operations Research*, 56 (1), 48-58.

Cullmann, A., 2012. Benchmarking and firm heterogeneity: A latent class analysis for German electricity distribution companies. *Empirical Economics*, 42 (1), 147-169.

Greene, W.H., 2008. The econometric approach to efficiency analysis, in: Fried, H.O., Lovell, C.A.K., Schmidt, S.S., (eds), The Measurement of Productive Efficiency and Productivity Growth, Oxford University Press, New York.

Jamasb, T., Nillesen, P., Pollitt, M. 2003. Gaming the regulator: A survey. *Electricity Journal*, 16 (10), 68-80.

Jamasb, T., Nillesen, P., Pollitt, M., 2004. Strategic behaviour under regulatory benchmarking. *Energy Economics*, 26, 825-843.

Johnson, A.L., Kuosmanen, T., 2011. One-stage estimation of the effects of operational conditions and practices on productive performance: Asymptotically normal and efficient, root-n consistent StoNED method. *Journal of Productivity Analysis*, 36, 219-230.

Johnson, A.L., Kuosmanen, T., 2012. One-stage and two-stage DEA estimation of the effects of contextual variables. *European Journal of Operational Research*, 220 (2), 559-570.

Jondrow, J., Lovell, C.A.K., Materov, I.S., Schmidt, P., 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics*, 19 (2-3), 233-238.

Kumbhakar, S.C., Lovell, C.A.K., 2000. Stochastic frontier analysis. Cambridge University Press, New York.

Kuosmanen, T. 2008., Representation Theorem for Convex Nonparametric Least Squares. *The Econometrics Journal*, 11, 308-325.

Kuosmanen, T., 2012. Stochastic semi-nonparametric frontier estimation of electricity distribution networks: Application of the StoNED method in the Finnish regulatory model. *Energy Economics*, 34 (6), 2189-2199.

Kuosmanen, T., Johnson, A.L., 2010. Data envelopment analysis as nonparametric least squares regression. *Operations Research*, 58 (1), 149-160.

Kuosmanen, T., Kortelainen, M., 2012. Stochastic non-smooth envelopment of data: Semi-parametric frontier estimation subject to shape constraints. *Journal of Productivity Analysis*, 38 (1), 11-28.

Lee, C.Y., Johnson, A., Moreno-Centeno, E., Kuosmanen, T., 2013. A more efficient algorithm for convex nonparametric least squares, European Journal of Operational Research, in press, available at: http://dx.doi.org/10.1016/j.ejor.2012.11.054.

Olson, J.A., Schmidt, P., and Waldman, D.M., 1980. A Monte Carlo study of estimators of stochastic frontier production functions. *Journal of Econometrics*, 13, 67-82.

Simar, L., Wilson, P.W., 2007. Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics*, 136 (1), 31-64.

Simar, L., Wilson, P.W., 2008. Statistical inference in nonparametric frontier models: recent developments and perspectives. In Fried HO, Lovell CAK, and Schmidt SS (eds.) Measurement of productive efficiency and productivity growth, 2nd edition, Oxford University Press.

Syrjänen, M., Bogetoft, P., Agrell, P., 2006. Analogous efficiency measurement model based on stochastic frontier analysis. Gaia Consulting Oy 11.12.2006 (Available at: www.energiamarkkinavirasto.fi).

Article 3

Antti Saastamoinen. Heteroscedasticity or Production Risk? A Synthetic View. Forthcoming in Journal of Economic Surveys, published online. DOI: 10.1111/joes.12054

JOURNAL OF ECONOMIC SURVEYS

doi: 10.1111/joes.12054

HETEROSCEDASTICITY OR PRODUCTION RISK? A SYNTHETIC VIEW

Antti Saastamoinen

Department of Information and Service Economy
Aalto University School of Business

Abstract. Two veins of literature, namely, production risk literature and stochastic frontier analysis, are examined. Both fields are concerned of output variation; the former due to exogenous shocks, the latter due inefficiency. By covering the literature from both the fields, this review suggests that the concept of heteroscedasticity can be utilized to build a synthesis between these mainly separate branches of literature. However, the synthetic approach brings a challenge how to differentiate between different sources of output variation. This challenge is identified as the main obstacle to meaningfully combine the two approaches.

Keywords. Heteroscedasticity; Just and Pope model; Production risk; Stochastic frontier

1. Introduction

In economics, the standard definition of production function defines it as the function that gives maximal output as a function of the given inputs (e.g. Varian, 1992). This definition implies that a producer is operating efficiently and is not facing any exogenous shocks to its input–output correspondence. Obviously, this situation rarely occurs in reality. The standard approach in the econometric estimation of production function has been to augment the otherwise deterministic input–output relation with a stochastic random error. Thus, the estimated production function does not need to correspond exactly to the observed production. Usually the random error is included purely for statistical reasons and the interest is in estimating the production function itself. However, two strands of production economics, namely, the production risk field (see, e.g. Moschini and Hennessy, 2001, pp. 110–112) and the frontier field (see surveys by Murillo-Zamorano, 2004; Fried *et al*., 2008), have positioned themselves to study the variation of production in terms of this error.[1]

The study of production risk has been prevalent in agricultural economics since uncertainty over output is especially present in agriculture, which is characterized by uncertain production environment due to, for example, weather, pests and pollution (Just and Pope, 2001, pp. 643–647; Moschini and Hennessy, 2001).[2] Such risk/uncertainty is often labelled as *production* or *output risk*.[3] Obviously, production risks also occur in other fields of production. However, for example, in manufacturing the consistent factory-like production environment significantly decreases the occurrence of such risks. In some sense, the production risk in agriculture resembles the well-known definition of risk in traditional finance literature where risk can be seen as the volatility (variance) of return associated with a given investment portfolio. In a similar manner we could assume that a farmer faces a certain degree of risk given the input factor portfolio.

The frontier field however assigns variations in output between producers mainly to the inefficiency of the producers. Inefficiency is measured against an ideal best-practice frontier. Moreover inefficiency is generally considered to be due to actions that are under the control of producer. Thus, the view of the

Table 1. The Positioning of the Review within the Literature.

	Regression	Frontier
Heteroscedasticity	Harvey (1976) Amemiya (1977)	Caudill *et al*. (1995) Hadri (1999)
Production risk	Just and Pope (1978, 1979)	Kumbhakar (1993) Battese *et al*. (1997) Kumbhakar (2002a)

frontier field seems distinctively different to that of the production risk field, which considers exogenous output shocks. However, varying operating conditions make efficiency measurement challenging as comparing the efficiency of producers makes sense only if the producers can be assumed to be operating in a relatively similar environment. Otherwise, it could be the case that some producers are seen as being more efficient only because they operate in a more favourable environment. This problem has long been acknowledged within the frontier field and consequently solutions have emerged to account for production environment. The core idea of more well known solutions has been to assume that the expected inefficiency or variance of inefficiency is producer- or environment-specific. In terms of econometric jargon, inefficiency is heteroscedastic in the latter case. The following example illustrates this point.

Consider producers in regions A and B. In region A, the variability of performance is higher than in region B. Assume also that the conditions of production environment in region A are more volatile than in region B. Now consider why the variation of performance might be higher in region A. We can assume that producers in either region differ in their capabilities to adapt themselves to the changes in their production environment. Thus, initially significant variation in performance can be present in region B also. But it is likely that in the long run, the more stable operation conditions lead to similar and predictable responses in this region. In region A on the other hand, the performance variations can persist simply due to a more risky environment. In this context, the risks of production manifest themselves as performance or inefficiency variations. Using again an analogy drawn from finance; both returns and losses are likely to be higher under high volatility. Evidently the riskiness of environment affects the production performance. Thus, it is surprising to notice that the concept of production risk is rarely mentioned in the frontier literature although heteroscedasticity is often discussed. This is despite that the production risk and the frontier fields share a substantial common ground in empirical applications in agriculture (see surveys by Battese, 1992; Bravo-Ureta *et al*., 2007).

Probably due to the profound conceptual differences between the two fields, historically neither field has been very aware of the other. Thus, a systematic and simultaneous exploration of the two fields would be warranted to see whether the fields have more in common than what the historical retrospective might suggests. Unfortunately, to our knowledge no such examination has been conducted. This review aims to fill this gap. We explore how ideas from both the fields have contributed to a synthesis in knowledge. This idea of synthesis we base on the concept of heteroscedasticity. Conceptually this review is positioned as in Table 1. There we have roughly identified four branches of literature. The literature in the upper left-hand cell refers to the traditional literature on heteroscedasticity in regression analysis. It serves as reference point for the three other branches that we deal in this review. The synthetic literature, which seeks to combine ideas from the two fields discussed, is located at the lower right-hand cell.

Our focus is on the empirical literature as the estimation of heteroscedastic econometric models has been a most fruitful area of convergence between the fields. The review adopts a microminded approach instead of a meta-analysis. We do not provide a full coverage of the literature in production risk, frontiers, heteroscedasticity or in agricultural economics – these goals have already been achieved

elsewhere. Agricultural production is however an integral part of the discussion because of the common ground of applications. Besides an expositional coverage of the subject, some critical insights regarding the possible challenges in combining the two fields are presented. This discussion also reaches beyond the core topic of heteroscedasticity. The synthetic view of the two fields constructed in this review should benefit the future research in bridging the gap between the fields. The examination in this study proceeds such that Sections 2 and 3 cover the production risk literature and frontier literature separately. Section 4 examines the studies, which aim to synthesize the ideas from these fields. Section 5 discusses the potential challenges, such as production dynamics and heterogeneity, of synthesis. Finally Section 6 concludes.

2. A Brief Introduction to Production Risk

Risk is not a new concept in agricultural production economics. However for our purposes, the appropriate origin of the literature dates back to the 1960s. Then the emphasis was on studying the distributional characteristics of crop yields by examining the moments of yield distribution. This also provided information about how input use might affect these moments. In his influential article, Day (1965) proposed that crop yield distributions ought to be positively skewed instead of being symmetrically normal. Risky weather conditions should imply less than mean yields as being the most likely. Only under ideal weather conditions, could extraordinary yields be obtained. For Day, the skewed distribution was a sign of risk. He also considered how different levels of fertilizer (nitrogen) use would affect the skewness of a crop yield. Day concluded that generally, a higher level of nitrogen use '. . . *places him* [farmer] *in a more favorable risk environment*'. implying that the risk is reduced by the use of nitrogen. This was in contrast with Fuller (1965) who found that the variability of yields increased along with nitrogen use (see also Just and Pope, 1979). Fuller, however, targeted his attention to variance instead of skewness as a measure of risk (see also Anderson, 1973).[4] These early analyses of risk in agriculture were much grounded on the distributional analysis of crop yields. From the perspective of production economics, the more production function-minded work of Just and Pope (1978, 1979) is generally considered to be the starting point of production risk literature in its current form.

Even today, the work of Just and Pope (1978, 1979) is often regarded as one of the hallmark models of production risk. Just and Pope (1978) criticized the traditional stochastic input–output responses such as Cobb–Douglas production function, since they impose strict constraints on how inputs affect the observed output variance, that is, the production risk. All traditional production functions applied held inputs to have only a risk-increasing effect on the output variance. Instead, Just and Pope suggested formulating the production function as shown in equation (1):[5]

$$y = f(X) + h(X)\varepsilon \tag{1}$$

The Just–Pope model (JP-model hereafter) in equation (1) combines the deterministic production function $f(.)$, which is a function of inputs X, with the additive stochastic error term ε with zero mean and constant variance. However, the term ε is scaled by a risk function $h(.)$. Just and Pope (1978) show that this formulation does not restrict the sign of the marginal risk effects of inputs on the variance of output. They also present a consistent and asymptotically efficient maximum likelihood (ML) estimation procedure to estimate the parameters in functions $f(X)$ and $h(X)$. The follow-up study by Just and Pope (1979) presented a three-step feasible generalized least squares (FGLS) estimator for obtaining the parameter estimates of the risk function. This latter estimator has subsequently been the tool mostly applied in empirical work (see Saha *et al*., 1997). Both estimation approaches take much from the traditional estimation of heteroscedastic regression models such as Harvey (1976) and Amemiya (1977) as the JP-model is in fact a model with heteroscedastic errors due to function $h(.)$. It is also important to note that the JP-model considers producers operating efficiently. Consequently, it assigns all variation from the optimal production levels only to exogenous shocks.

The JP-model framework has been fairly popular in empirical applications (e.g. Traxler *et al*., 1995; Kumbhakar and Tveterås 2003). Asche and Tveterås (1999) employ the JP-framework to estimate the risk effects in the Norwegian salmon farming industry. Departing from the FGLS procedure presented by Just and Pope, they estimate the production function and the risk function with ordinary least squares (OLS) in two separate steps, exploiting the fact that production uncertainty can be treated as heteroscedasticity. The approach by Asche and Tveterås avoids using non-linear least squares needed in FGLS approach but they lose some of its estimation efficiency. They also discuss the role of heteroscedasticity in risk considerations. They see heteroscedasticity as a phenomenon with some economic content instead of purely econometric misspecification problem. This review concurs with this interpretation. Consequently they propose that typical heteroscedasticity tests can be applied to detect production risk. Of course, these tests may also pick up misspecification that is not production uncertainty. Another novel application of the JP-framework by Koundouri and Nauges (2005) augments the JP-model with a Heckman-type selection model. According to them, the parameters of the risk function can be biased if crop selection is not taken into account. It is reasonable to expect that risk parameters can be affected by crop selection as selection is often determined by the same variables (e.g. weather) that constitute the variability in output.

Relatively quickly after its introduction, the JP-model was noted by Antle (1983b) as still rather restrictive. According to Antle, the JP-model imposed similar restrictions on higher order moments (above the second moment) than the models criticized by Just and Pope (1978) imposed on the second moment. Antle showed that the elasticity with respect to input(s) of any higher order moment is directly proportional to the elasticity of the second moment.[6] For Just and Pope, this was not a problem as they considered only variance as the relevant measure of output risk. Thus, the research line initiated by Antle revived the discussion seen already between Day and Fuller about the proper measures of risk. Antle also proposes a flexible moment-based estimation approach where the effects of inputs on each moment are allowed to differ.

Antle and Goodger (1984) applied Antle's earlier framework in a milk production application. They found that after accounting for the third moment, the input use of the risk-averse decision maker may change when compared to the traditional mean-variance set up. Antle (1983b) and Antle and Goodger (1984) also contemplated the relevant number of moments. Following Kendall and Stuart (1977), Antle (1983b) viewed moments up to the fourth moment as being sufficient to describe the yield distribution. Of course, the number of relevant moments is ultimately an empirical matter. Recently, Antle (2010a) suggests that partial moments instead of full moments provide a more detailed description of inputs' effects on moments since inputs are allowed to have different effects on the moments in the different parts of the output distribution. Further discussion can be found, for example, from Du *et al*. (2012) who extend the JP-model to include also skewness.

With a slight digression, we note that we have not yet touched upon the issue of risk preferences of the producers. For example, Love and Buccola (1991) call for a joint estimation of technology and risk preferences since technology parameters are inconsistent if these preferences are not accounted for. Lence (2009) and Just and Just (2011) have, however, pointed out some challenges in estimating risk preferences from the production data. Preferences are also significant when comparing alternative technologies. Tveterås (1999) notes that in a deterministic setting, technical efficiency is a sufficient condition to rank two alternate technologies. But as the setting becomes risky (stochastic), efficiency is no longer an objective measure, as the ranking of technologies depends on the risk preferences of the producer. In other words, the ranking of technologies depends on how the producer prefers the expected output over its variance. Since the review of risk preference literature would practically merit a review study of its own, we do not diverge further on this topic.[7]

It seems that production risk literature has been divided between the variance-based and the skewness-based measurement of risk. The core question is whether the first two moments are an adequate representation of output distribution and its riskiness.[8] Just and Weninger (1999) suggest that crop yields

can be adequately represented with a normal distribution. Antle (2010b) argues that it is an empirical and testable question as to whether, for example, any two parameter location-scale distribution is adequate. He points out that location-scale assumption is at the heart of the JP-model with heteroscedastic errors. But considering the next section on heteroscedastic frontier models, we view that the relevant comparison point for these models is the JP-model due to its direct grounding on heteroscedastic regression models. It is, however, important to be aware of the basic higher order moment-based approaches of estimating production risk, since skewness of the error term is one of the central issues of the frontier field also (see, e.g. Kuosmanen and Fosgerau, 2009 and the next section). It is especially interesting that skewness is given different meanings in these fields; a point that will become evident in our future discussion.

3. Frontier Literature on Heteroscedasticity

We start with a very brief overview of the frontier field for those not familiar of it. The field targets an estimation of a frontier, which usually defines maximal production. Different from typical production function estimation, the producers are not assumed to operate efficiently. Inefficiency is then a one-sided deviation from this frontier. This also suggests a different error structure for estimation than in a typical production function with random symmetric noise. The general definition of technical efficiency by Koopmans (1951) states that the firm is technically efficient if any increase in one output would imply a reduction in at least one other output or increase at least one input, The origins of efficiency measures are by Debreu (1951) and especially by Farrell (1957), whose definitions of the measures are based on the radial expansion of outputs or the radial contraction of inputs. The development of current efficiency estimation methods however dates back to end of 1970s, interestingly coinciding with the development of the JP-model.

The frontier field has been characterized by two efficiency estimation paradigms. The non-parametric data envelopment analysis (DEA) concept was originally suggested by Farrell (1957) but popularized by Charnes *et al*. (1978). Since, it has been the predominant approach within the operation research and management science community. DEA has gained its popularity mainly due to its axiomatic approach in defining the efficient production frontier. This approach relies on a minimal set of regularity conditions for production technology such as monotonicity, concavity and certain returns-to-scale. Given these assumptions, the technology is estimated with a mathematical linear programming optimization. The main limitation of DEA is that it assumes all deviation to the frontier being due to inefficiency. The inability of DEA to account for statistical noise and measurement errors gave rise to another strand of frontier estimation with a strong econometric background. Stochastic frontier analysis (SFA) (Aigner *et al*., 1977; Meeusen and van den Broeck, 1977) incorporates statistical noise into its estimation framework. It however imposes a functional form for the production technology. Typically it also requires assumptions regarding the distribution of inefficiency and noise. Due to the econometric groundings of SFA, it has achieved a more fruitful treatment of heteroscedasticity. Thus, we limit our interest on SFA literature and leave an exploration of DEA and other semi- or non-parametric methods under heteroscedasticity as a topic of further research. Many semi- and non-parametric methods are anyhow more robust with respect to heteroscedasticity than SFA as they attempt to relax some of the assumptions of the parametric models (see, e.g. Kumbhakar *et al*., 2007). These approaches however pay little attention to model heteroscedasticity as an economic phenomenon.

The original SFA model that Aigner *et al*. (1977) proposed for the estimation of production function in the presence of noise and inefficiency is in equation (2):

$$y_i = f(\mathbf{x}_i; \beta) + \varepsilon_i \tag{2}$$

where $\varepsilon_i = v_i - u_i$.

In equation (2), the deterministic part of the production frontier $f(\mathbf{x}_i; \beta)$ specifies the maximum production given inputs $\mathbf{x}_i$ for firm i. The parameter vector β represents the parameters of the production

function. Some parametric functional form (e.g. Cobb–Douglas, translog) is assumed for $f(.)$. The composed error term, ε_i represents the overall deviations from the frontier. The term, v_i is a typical two-sided symmetric noise term. Consequently $f(\mathbf{x}_i;\beta)+v_i$ forms the stochastic frontier. Aigner *et al*. (1977) characterize v_i being the result of external events outside the control of a firm such as luck, weather and topology. Another source for v_i is in specification and measurement errors. The term, u_i is the one-sided inefficiency component. For production function, it is always assumed as positive. The parameters β are typically estimated with ML although OLS-based approaches are also available. Using ML necessitates some distributional assumptions for the error components. The usual assumptions are $v_i \sim N(0,\sigma_v^2)$ and $u_i$that follows some one-sided distribution such as half-normal, $u_i \sim |N(0,\sigma_u^2)|$ or exponential. It is also assumed that u_i and v_i are independent of each other and of inputs $\mathbf{x}_i$.

One complication in the early stages of SFA literature was that in a cross-sectional case, only industry-wide average inefficiency could be obtained. Jondrow *et al*. (1982) derived the point estimator for firm-specific inefficiency as $E(u_i|\varepsilon_i)$ and this estimator still frequently applied despite being inconsistent in a cross-sectional case. Since the distribution of inefficiency is one-sided, the composite error ε_i should be negatively skewed in the case of a production frontier. Thus, a test on the skewness of $\hat{\varepsilon}_i$ serves as a specification test, as to whether the frontier formulation is appropriate in contrast to a neoclassical production function (Kuosmanen and Fosgerau, 2009).[9]

It is also reasonable to ask whether distributional assumptions on inefficiency have any major effect on the SFA estimation. The choice of the distribution for the inefficiency term is relatively arbitrary as long as the chosen distribution is one-sided. Generally, the choice does not significantly affect the estimated efficiency scores or their rankings (e.g. Greene, 2008, pp. 180–184). Certain distributional assumptions however may have some interesting conceptual implications as we later note. Relaxation of these assumptions is possible by using panel data or generalized method of moments (GMM) estimation (Schmidt and Sickles, 1984; Kopp and Mullahy, 1990).

Heteroscedasticity is introduced by allowing the variances of the error components to be observation-specific, that is, $\sigma_{v,i}^2$ and/or $\sigma_{u,i}^2$. Within frontier literature, the discussion of heteroscedasticity is necessarily preceded by a more general discussion of the modelling of exogenous inefficiency effects with so-called z-variables (see survey in Kumbhakar and Lovell, 2000, Chapter 7). The z-variables are usually considered to be variables which are not part of the production technology as such but which can still affect the efficiency of the production process. These variables thus affect the relative location of the frontier but not the shape of it. These may include variables describing, for example, the operating environment or producer-specific characteristics. The early and rather intuitive approach to assess the effects of z-variables was to first obtain a measure of efficiency and then run a regression of the obtained efficiency scores on z-variables. This approach has since been proven to be unsatisfactory due the complications concerning the two separate steps of estimation (see, e.g. Schmidt, 2010). Thus, usually the z-variables parametrize the distribution of inefficiency and the effects of them and the production function are jointly estimated in a single stage.

Kumbhakar *et al*. (1991) (KGM hereafter) parametrizes the mean of inefficiency distribution such that $u_i \sim |N(\mathbf{z}_i\gamma,\sigma_u^2)|$. The parameter vector γ gives the effects of z-variables on mean inefficiency. This approach however confounds the effects of z-variables on the expected level of inefficiency and the variance of inefficiency. Shifting the underlying mean of the untruncated u necessarily shapes the variance of the truncated u. Extensions of the KGM model are Huang and Liu (1994) and Battese and Coelli (1995). Huang and Liu's model (1994) includes the interactions of inputs and z-variables in the vector $\mathbf{z}$. Thus, the inefficiency effects may vary according the input level. Battese and Coelli (1995) extend the model to panel data. By parametrizing the expected inefficiency, these models shift the relative location of the frontier. The location of the frontier can also be shifted by other means. Nothing in our earlier definition of z-variables prevents them being entered into the model as in equation (3), as suggested

by Reifschneider and Stevenson (1991):

$$y_i = f(\mathbf{x}_i; \beta) + g(\mathbf{z_i}; \delta) + \varepsilon_i \tag{3}$$

In equation (3) the z-variables act as direct production function shifters. The practical effects of KGM (1991) and Reifschneider and Stevenson (1991) specifications are quite similar. As a consequence, it is difficult to differentiate between these two positioning (see discussion in Fried *et al*., 2008, pp. 156–157). Conceptually the interpretations, however, will differ. In the latter case, the production function is shifted with respect to the observations, whereas in the former case the observations are effectively adjusted by their environment/characteristics with respect to the frontier.

None of the above studies explicitly model heteroscedasticity with z-variables. Reifschneider and Stevenson (1991) noted that the standard deviation of inefficiency could be modelled as a non-negative function of the z-variables but did not implement this approach. With simulation experiments, Caudill and Ford (1993) and Caudill *et al*. (1995) (CFG hereafter) showed that unaccounted heteroscedasticity biases the parameter estimates of frontier function and the obtained efficiencies. To account for heteroscedasticity, CFG (1995) suggested the parametrization shown in equation (4) for the standard deviation of inefficiency. In the equation, $\mathbf{Z}$ is a vector of variables mainly related to firm size (their definition) and γ is again a vector of the unknown coefficients to be estimated. The exponential form for parametrization guarantees that the standard deviation is non-negative:

$$\sigma_{ui} = \exp(\mathbf{Z_i}\gamma) \tag{4}$$

$$\sigma_{vi} = \exp(\mathbf{W_i}\theta) \tag{5}$$

Subsequently, Hadri (1999) proposed an obvious extension to the CFG (1995) model with his *doubly heteroscedastic model*, by parametrizing also the standard deviation of the two-side noise component within the same model. Thus, Hadri's model encompasses the CFG (1995) model as a special case if the standard deviation of noise is constant. Notice that parametrizing only the standard deviations or variances of the distributions does not necessarily offer much more clarity compared to the above models, which parametrize the mean only. If the underlying variance is parametrized, keeping the pre-truncated mean constant, then the after truncation mean is again necessarily changed. It also serves to briefly discuss the variable choice at this point. Hadri viewed the Z-variables in equation (4) to be variables related to firm management, whereas the W-variables in equation (5), he viewed as size related variables. In the context of this review, it is however more relevant to ask whether a variable is under or beyond the control of firm management. It would be tempting to say that Z-variables are more in the control of management than W-variables as unlike noise, inefficiency should be controllable. However, while noise itself is not controllable, the effects of it may be, at least considering from a production risk perspective. Thus, is not directly obvious that all W-variables are uncontrollable as in the spirit of JP-model it could include inputs, for example.

Concluding this section, the most pressing issue in modelling exogenous efficiency effects is well characterized by Greene (2008, p. 154) when he asks '*Where do we put the z's*'? The problem is, that regardless of the placement of the z-variables their practical effect is often much the same and thus hard to explicitly identify. For modelling risk, however, the more important issue is whether the standard deviations given in equations (4) and (5) can be interpreted as production risk or uncertainty. The discussion in the following section will endeavour to show that indeed they can be given such meaning.

4. Convergence of the Two Fields: The Synthesis

In this section, we examine how the ideas from the production risk and frontier fields have been synthesized. We begin with a statement from Gallagher (1987): '*Capacity is defined here as the yield that would occur with efficient use of the given technology for controllable inputs and ideal weather*'. This

statement well captures the importance of the operating environment in frontier models. More specifically it says that the maximum capacity is achieved only under ideal conditions. With capacity Gallagher referred to a frontier and according to him the deviations from this frontier are due to one-sided random fluctuations, that is, production risks. Thus, Gallagher does not have any inefficiency considerations in his study. That is why we do not consider the model by Gallagher as a synthetic approach in the sense what is meant in this review as it does not discuss inefficiency and production risk together.

Somewhat closer to such synthesis is the work by Antle and Crissman (1990) who derive a measure of technical efficiency within an expected utility framework. Their efficiency measurement concerns the differences in efficiency between traditional and modern crop varieties. The relative efficiency between the producers of these varieties is the ratio between their expected utilities, which again depend upon the producer surplus. The moments of producer surplus are functions of inputs. Thus, the uncertainty over output is taken into account in the producer problem by defining the objective as an expected utility of the producer surplus instead of surplus itself. Since their efficiency comparisons are only pairwise, they do not estimate any frontiers and thus again their work cannot be considered as synthesis quite in the sense of this review. Their work, however, raises an important issue of production dynamics and its implications on efficiency measurement (see Section 5).

Closer to the typical SFA framework, Kumbhakar (1993) stated the need for a joint estimation of technology, production risk and inefficiency. He views that incorporating risk into to the basic SFA framework makes SFA model heteroscedastic. Indeed, from equation (6), we see that his model is an extension of the JP-model. The actual estimation much follows the panel data model of Griffiths and Anderson (1982). Griffiths and Anderson present a framework for estimating firm-specific effects under the Just and Pope risk specification without considering these individual effects as inefficiency. Kumbhakar gives them such an interpretation. Unfortunately, also the main challenges of Kumbhakar's model relates to these firm-specific effects. First they are assumed to be fixed over time. It may not be reasonable to assume this in longer panels. The model nevertheless allows these firm-specific effects on output to vary along with the input use in Just–Pope fashion, but the underlying inefficiency is fixed. Secondly, as we will later discuss, the fixed effects component faces problems when we attempt to identify it as inefficiency, as it easily picks also other heterogeneity that is not related to inefficiency:

$$\ln y = \ln f(x;\alpha) + g(x;\beta)[\tau_i + \lambda_t + v_{it}] \tag{6}$$

where

τ_i is the time-invariant firm effect/inefficiency
λ_t is the time-specific effect,
v_{it} is the firm- and time-specific random noise and
α, β are the parameters associated with the production and risk functions.

Strictly in SFA context, Battese *et al*. (1997) (BRW) introduced a cross-sectional stochastic frontier model with a composed error term and flexible risk properties as shown in equation (7). All terms in equation (7) are similarly defined as before apart from the non-negative u_i which is now distributed as the truncation of the $N(\mu, \sigma_u^2)$ distribution. We see that the standard SFA and JP models result as special cases of the model in equation (7) if we omit the function $g(.)$ or inefficiency correspondingly. In their application the risk flexible model or the traditional JP-model did not yield noticeably different marginal products of inputs, compared to the typical SFA model. Moreover, inefficiency effects were tested being absent in stochastic flexible risk model, suggesting that it does not differ significantly from a typical JP-model. Finally, the parameter estimates of the risk function did not differ between the JP-model and the risk flexible SFA specification. The last finding is quite expected as BRW do not identify which error component is the source of risk since the function $g(.)$ is the same for both v and u:

$$Y_i = f(\mathbf{x}_i;\alpha) + g(\mathbf{x}_i;\beta)[v_i - u_i] \tag{7}$$

The natural extension of the BRW (1997) model was presented by Kumbhakar (2002a). This model is shown in equation (8). Kumbhakar aims to augment the SFA model with production risk and on the other hand to account for the possibility of inefficiency in the JP-model. Most importantly, however, Kumbhakar also includes the estimation of risk preferences in this model. Thus, the model allows us to elicit information as to how both technical inefficiency and production risk contribute to the input decisions given the risk preferences. Neither the traditional JP-framework nor the SFA-framework account for preferences, so preferences are not assumed to affect the input use. In reality a risk-averse producer could prefer risk-reducing inputs over other inputs.

$$y = f(\mathbf{x}, \mathbf{z}) + g(\mathbf{x}, \mathbf{z})v - q(\mathbf{x}, \mathbf{z})u \tag{8}$$

Note that, Kumbhakar defines the vector $\mathbf{z}$ as quasi-fixed inputs, not as the usual z-variables we have discussed earlier. It is straightforward to see that the BRW-model results as a special case if $g(\mathbf{x}, \mathbf{z}) = q(\mathbf{x}, \mathbf{z})$ and the standard JP-form is obtained if no inefficiency is present. Functional form for all the functions $f(.)$, $g(.)$ and $q(.)$ has to be specified. It is often argued in favour of semi- or non-parametric frontier models that any pre-determined functional form is hard to justify for the production function. Just and Weninger (1999) pointed out that functional form misspecification is often a major source of complication in the estimation of crop yield distributions. Thus, it might be hard to justify any specific form for the functions $g(.)$ and $q(.)$ also. Kumbhakar and Tsionas (2010) partly relax these parametric restrictions and consider non-parametric kernel estimation of the production and risk functions. They however remain within the standard JP-model and do not consider inefficiency to be present. Relaxing these parametric assumptions further in this context might provide a potential avenue for further research.

Kumbhakar applies his model to data on Norwegian salmon farmers. He finds that for risk-averse farmers, the output risk impacts more on their input decisions than technical inefficiency. This implies that the risk-averse producer is more concerned in reducing the output risk at the expense of the efficient use of inputs. Already Ramaswami (1992) pointed out that the marginal risk premium is negative for risk averters if and only if the input is risk decreasing. In contrast, a risk-averse producer should be compensated (positive risk premium) for the use of risk-increasing input. Thus, in sectors where risk considerations are of concern but where firms are risk neutral, the firms may not have sufficient incentives to reduce risk. Consider, for example, the regulation of public utilities and their service provision. Often incentive schemes in regulation emphasize efficiency over quality factors (see, e.g. Giannakis *et al*., 2005). Furthermore, the minimization of inefficiency and risk are often contradictory objectives as the high use of risk-reducing input can appear as technical inefficiency (see Kumbhakar, 2002a). Therefore, the regulated utilities may be more concerned with being technically efficient, rather than improving the quality (riskiness) of their services.

The models by Kumbhakar (1993, 2002a) and BRW (1997) are JP-augmentations of the standard SFA model. Nevertheless, the concept of production uncertainty has been introduced in SFA literature without any reference to the JP-framework. Instead of the parameters of risk function, Bera and Sharma (1999) targeted their analysis towards the variability of inefficiency. Following the Jondrow *et al*. (1982) estimate of inefficiency $E(u_i|\varepsilon_i)$, they introduce the conditional variance of inefficiency $\text{Var}(u_i|\varepsilon_i)$ as a measure of production uncertainty, where ε_i is the composed error term from a typical SFA model.

Bera and Sharma show that both $E(u_i|\varepsilon_i)$ and $\text{Var}(u_i|\varepsilon_i)$ are monotonically decreasing in ε_i for production function. This implies that given a fixed v_i, the closer the production is to the frontier, the less production uncertainty it faces. They also note that the possibility for efficiency improvement is larger under a high uncertainty. That is, the largest scope for efficiency improvement is where the largest variability of efficiency occurs. However, we must take some care in interpreting $\text{Var}(u_i|\varepsilon_i)$ as a measure of production uncertainty. If we compare the result of Bera and Sharma to Kumbhakar (2002a), they seem somewhat contradictory. In Kumbhakar's approach, the objectives of technical efficiency and production uncertainty can be in contradiction due to risk aversion. In Bera and Sharma's approach, these measures go to the same direction. Bera and Sharma acknowledge that production uncertainty can be due to factors

other than inefficiency (e.g. environment). There is no need to assume that these factors are less volatile near the frontier. In fact, the producer may need to be more efficient in order to successfully operate in a riskier environment. Thus, the most efficient units may face the highest production uncertainty. Finally, note that, due to conditioning on ε_i, the noise component is fixed in their approach. Thus, factors in v cannot be considered to be contributing to production risk. For future research they propose to investigate the conditional skewness and kurtosis measures (see also Asche and Tveterås, 1999, and the reference therein).

Wang (2002) also considers the variance of inefficiency as the measure of production uncertainty. For analytical simplicity, Wang uses the unconditional variance of inefficiency $\mathrm{Var}(u_i)$ instead of $\mathrm{Var}(u_i|\varepsilon_i)$. Wang however places more interest on how z-variables affect his measure of production uncertainty. We have previously discussed some developments of modelling the exogenous efficiency effects in SFA models. Those models were mainly concerned with parametrizing either the mean or the variance of inefficiency distribution. Wang (2002) proposes a model where the z-variables affect both the mean and the variance such that $u_{it} \sim N^+(\mu_{it}, \sigma_{it}^2)$. The detailed parametrizations are shown in equations (9) and (10).

$$\mu_{it} = \mathbf{z}_{it}\delta \tag{9}$$

$$\sigma_{it}^2 = \exp(\mathbf{z}_{it}\gamma) \tag{10}$$

This model seems to be a direct extension of the models by KGM (1991) and CFG (1995). The model allows the z-variables to have non-monotonic marginal effects on inefficiency within the sample. This means that the effect of z-variables on $E(u_i)$ and $\mathrm{Var}(u_i)$ can differ in magnitude and even in sign with different values of z-variables. Wang notes that in KGM framework, a certain z-variable is either efficiency enhancing or efficiency impeding. To illustrate his model, Wang uses farmer's age as an example. For younger farmers, ageing affects positively on productivity through gaining more experience. Above a certain age the effect turns to be negative as older farmers are physically less capable to perform well. The effects on variance are generally the same. Wang however notes that in his model, the effects on mean and variance can also differ. Nonetheless, it is not directly observable whether the Wang model suffers from the same problem of confounding effects as the KGM (1991) model. Moreover, it is not obvious either from the work of Bera and Sharma or the one by Wang, whether their use of the term 'production uncertainty' holds significantly different connotations to what is meant by production risk in the sense of Just and Pope. But since variations in inefficiency arguably translate to variations in output, these concepts can be seen to coincide at some level.

In a parallel study to Wang (2002), Wang and Schmidt (2002) propose a model shown in equation (11) with a so-called *scaling property* for inefficiency (see also Simar *et al*., 1994). In equation (11), the distribution u^* is independent of z-variables. This so-called base inefficiency is however scaled with a *scaling function* $h(.)$ where δ is a parameter vector associated with the z-variables. Following Alvarez *et al*. (2006), the scaling property can be interpreted such that the basic distribution of u^* reflects, for example, the natural managerial abilities of the manager which are not affected by any contextual factors:

$$u(\mathbf{z}, \delta) = h(\mathbf{z}, \delta)u^* \tag{11}$$

The scaling function, which can be a function of the operating environment, then effectively scales up or down the mean and the spread of these managerial capabilities. In other words, the extent that the natural skills of the manager are used effectively depends, for example, on the manager's schooling and operating environment. Interestingly, the scaling function could also be interpreted as a risk function. Since the scaling function determines the scale of the distribution for inefficiency, a larger scale can be considered as a riskier environment. In the JP-model also the effects of random shocks are in some sense scaled up or down by the risk function.

The main contributions for the synthesis between the stochastic frontier models and production risk models can be considered to be those listed above. However, few other studies warrant a mention. Jaenicke *et al*. (2003) compare different SFA models and the JP-model in a cotton cropping system application. They found that the ordering of different cropping systems with respect to their riskiness was much affected by the chosen method. However, they do not combine the JP- and SFA-model in the same fashion that has been done above by BRW (1997), for example. Their final 'new' model is in fact a variant of the Wang (2002) model with the exception that Jaenicke *et al*. parametrize the variance of noise instead of the variance of inefficiency. Thus, they interpret the variance of noise to represent production risk. Huang and Kao (2006) extend the inefficiency/risk estimation to a multi-output setting as according to them any single output model is unable to identify risk from technical efficiency. They argue that since observable output is affected by the production risk, these factors become undistinguishable from output inefficiency. Instead they propose to associate risk to only one of the outputs and regard the other outputs as riskless. It is uncertain as to what constitutes the rules in determining risky output. In the Huang and Kao (2006) banking example, their approach may seem plausible, but in agriculture the multi-output farmer is likely to face risks relating to each of the outputs. Their notion of identification however merits its place.

This section outlines two distinctive approaches to synthesis. The JP-augmentation of SFA models (Kumbhakar, 1993, 2002a; BRW, 1997) can be considered as the more direct approach, whereas 'the variance approach' (Bera and Sharma, 1999; Wang, 2002; Jaenicke *et al*., 2003) examines the production risk indirectly from the variance of inefficiency or noise. The former approach allows us to infer the effects of input use on production risk with the cost of more complex estimation. The latter approach could in principle be extended to examine the risk effects of inputs by obtaining a measure of production risk and regressing that on the inputs. However, this type of approach might yield similar statistical issues to those known in the two-step approach of z-variable modelling in frontier literature (Schmidt, 2010). It is often argued that the validity of a two-step approach relies upon z-variables being separable from the input–output space (see, e.g. Daraio and Simar, 2005). Similarly, it can be argued that risk cannot be separated from technology estimation, especially if we study inputs' effects on risk. Thus, it may be preferable to jointly estimate the technology and risk parameters. This goal is partly achieved with the JP-augmented SFA-models. However, complications arise, as often such models are unable to adequately separate between inefficiency, production risk or any other form of heterogeneity.

5. Discussion and Further Extensions

As we have seen, the concept of heteroscedasticity appears to be the linking channel between the production risk and SFA models. Our treatment so far has largely been expositional. We have tangentially discussed some of the challenges of a synthetic approach but the implications of these challenges are yet to be studied in detail. In this section, we deal the challenges of a synthetic approach in a wider context and extend the discussion also beyond the concept of heteroscedasticity. There are two issues that we especially examine in this section. First, our previous discussion has not yet identified as to what stage of production process inefficiency and risk should be considered. But clearly production dynamics affects our perceptions of inefficiency and risk, as production decisions often are long-term decisions. Secondly, identification between inefficiency and production risk can be problematic since producer or environment heterogeneity may perturb our attempts to separate them.

The full exploration of production dynamics falls beyond the scope of this review. Dynamics, however, has some relatively intuitive implications for risk and inefficiency. So far, a relatively static view of risk has been presented. Up to this point, we have mainly considered risk to manifest itself in the uncertain output given the decision on inputs. But of course the riskiness of production may already affect the input decisions. Already Just (1974) found that past experiences of risk significantly affect the future decisions of a farmer. Implicitly the later Just and Pope (1978) model also aims to learn something about this feedback process as it aims to see how risk could be controlled with the input use. Jolly (1983) has

categorized the risk management of a farm to be compiled from two types of actions, namely, those actions against risk exposure and those in controlling risk impacts. The producer probably seeks to minimize both the exposure to risk and the impacts of a realized risk. In practice, the former may be impossible to control for but the latter seems more controllable through the use of risk-reducing inputs. Regardless of how controllable exposure and impacts are, the decisions concerning them are anyhow made with incomplete information. The farmer has only subjective *ex ante* evaluation regarding risk or uncertainty at hand at the time of the decision. Therefore, some degree of inefficiency may result from this informational deficiency. The risk-efficiency hypothesis by Antle (1983a) well summarizes the farmer's problem: '. . . *previously optimal decisions based on old information become suboptimal with new information. These facts lead me to hypothesize that risk affects both*, *productivity* (*technical efficiency*) *and optimal resource use* (*allocative efficiency*) . . . '. Simply put, previously technically (or allocatively) efficient decisions may not be efficient when subject to new information. Clearly, efficiency after resolved uncertainty is different from the efficiency before unresolved uncertainty. The problem for the analyst is that production analysis is often *ex post* analysis of the observed behaviour. Pope and Chavas (1994) and Pope and Just (1998) show that a so called *ex post* cost function conditioned on the observed output is not compatible with the expected utility maximization when the output is in fact stochastic. As a consequence, biased parameter estimates of the cost function are obtained if *ex post* function is applied instead of *ex ante* function.

Production dynamics also matter for technology adoption, which inherently is a dynamic process. As discussed by Antle and Crissman (1990), technology adoption has implications for efficiency measurement also. Consider that in a certain period, a farmer experiments with new technology to obtain possible future gains. The farmer may appear relatively more inefficient in this adoption period than in a previous period as the farmer is yet to fully master the new technology. Is it then correct to interpret this as inefficiency? Especially if the learning process results in significantly better outcomes in the future, a snapshot view on efficiency can be severely misleading. Thus, the long-term optimization problem might significantly differ from the short-term one as noted by Antle and Crissman (1990). Indeed they find that during the early periods following the introduction of a new technology, the adopters suffered efficiency loss relative to the users of old technology. This relationship is however reversed due to learning in later periods. Thus, in the present context we see that technology adoption can imply period-specific heteroscedasticity. Experimenters have higher variation in their output during experimental periods compared to the non-experimenters. Ghosh *et al*. (1994) also examine the role of technical inefficiency and risk attitudes in the technology adoption process and found that technically more efficient producers were more willing to adopt the new technology. They view the new technology as risky and as a result, the technically inefficient risk-averse producers are not willing to adopt the new technology. This is because in principle they can increase their expected profits with old technology by increasing their efficiency.[10] These results in together may explain why we may observe rather large efficiency variations of a producer between different periods.

The second complication of the synthetic treatment of risk and inefficiency comes from the identification between them. Identifying is challenging, as both are deviations from a production frontier. Here, we consider both deviations to be output reducing. Of course, we could have a positive shock, but in a case of risk it is maybe more natural to speak of negative shocks. More specifically, the problem is how to properly decompose the overall residual into these two (or even more) subcomponents. Before that, it is informative to ask a more general question about how we obtain this residual at first place. The residual is always a result of unaccounted factors of our model (or misspecification). Thus, the key is how much conditioning we can and want to do in our model. In practice we are forced to put some limit to our conditioning. By imposing this limit, something is necessarily left unexplained. From here, it is only a matter of labelling as to whether we call the remaining residual as inefficiency, production risk or noise. Abramovitz (1956) famously stated that the residual can be seen '*as the measure of our ignorance*'. Unfortunately, whether the residuals represent the ignorance of the firm under study or the analyst is often unclear. We can only hope that the analysts have the competence to include all the relevant factors

to minimize their ignorance. According to Stigler (1976), observed inefficiency might be a result of the failure of the model itself. [11] However, even after a correctly specified model it is not certain what does the obtained residual stand for. As said, this deviation can be manifested as inefficiency, production risk or noise. Thus, the next step would be to decompose the residual. The problem of decomposing the overall error to its parts (inefficiency and noise term) is known as the *deconvolution problem* in frontier literature (see, e.g. Amsler *et al*., 2009). The signal (inefficiency) that we are looking to extract is convoluted with the relatively uninteresting part (noise). Similarly, inefficiency can be convoluted with production risk (O'Donnell and Griffiths, 2006). Thus, what often is labelled as inefficiency might simply be a realized output risk. The deconvolution is further complicated as the convolution with production risk in fact involves three parts: inefficiency, production risk and also noise. O'Donnell and Griffiths (2006) propose to achieve this type of deconvolution in a Bayesian estimation framework.

If we are interested in both inefficiency and production risk within the standard SFA framework, the easy way out would be to assume that the noise term is production risk. Nevertheless noise has not usually been labelled as risk since the motivation to include the noise has been mostly statistical. Of course placing risk to the noise term raises the problem of how to differentiate risk from measurement errors and other statistical noise. By no means probably, but following the JP-framework analogy, there is where the risk component naturally falls. But the way we generally see statistical noise in an econometric model has an interesting consequence on how we assume risk being distributed if we consider noise as risk. Since the noise term is usually always assumed to follow a symmetric distribution, we would consequentially assume that production risk is symmetric. But already in Section 2, we noted that risk might have a skewed distribution. For example, small- and mediocre-sized risks are often more likely than high risks, thus leading to a positively skewed risk distribution. This however would complicate identification, as now both inefficiency and noise would be skewed similarly. Analogously, as Amsler *et al*. (2009) point out, nothing in principle rules out 'nearly normal' distributions for inefficiency. Again identification would be practically impossible. It may be because of these challenges that risk has more often been incorporated to the variance of the inefficiency term in frontier literature and the noise term is subsequently left without further interest.

Even more general problem is how to untangle inefficiency from producer and environment heterogeneity. Some authors in the efficiency literature are concerned that the variations in environment and/or the characteristics of producers are often misinterpreted as observed inefficiency. For example, O'Donnell *et al*. (2010) (OCQ hereafter) suggest that efficiency considerations with traditional frontier models may be misleading due their inability to take into account the uncertain production environment and the producers' views about this environment. OCQ suggest that information asymmetry between the producers can cause productivity differences even without any technical or allocative inefficiency. In other words, the productivity differences are not due to any inefficient use of resources, as we would normally understand inefficiency, but only due to differences in perceptions. Obviously the better informed producer is more capable to adapt himself to the uncertain possibilities of the future. OCQ (2010) illustrate their point in simulations where the expectations of risk among producers are heterogeneous.[12] They assume that producers are fully rational and optimize their production such that no inefficiency in its traditional sense is present. This assumption illustrates their main point that even among fully efficient producers, productivity varies due to the different states of nature. More specifically, they show that the traditional efficiency estimators, such as SFA and DEA, identify inefficiency being present although all producers are efficient. Thus, the unaccounted informational asymmetries are portrayed as inefficiency. They also assume that producers face the same set of possible states of nature. This rarely is the reality but this assumption highlights that producers differ only in their beliefs of future states, not in their possible states of nature.

The work by OCQ (2010) recognizes more or less the same message as the study by Greene (2004). Greene studied the differences in the efficiency of national healthcare systems using a large WHO data set. He points out that a substantial proportion of the country heterogeneity has been misinterpreted

as inefficiency in earlier studies on the subject. Greene notes that in typical panel data frontier models such as Kumbhakar (1993), the time-invariant unit-specific component practically 'masquerades' all heterogeneity as inefficiency. Instead Greene (see also Greene, 2005) suggests more general *true* fixed and random effects models, which add a further heterogeneity component to the typical panel data models. However, the true fixed effects model is often practically infeasible with a large cross-sectional dimension as the number of estimable parameters increases rapidly.

The above true fixed/random effect models do not consider heterogeneity in production function parameters. Greene (2005) actually shows that the above true random effects model is a special case of a more general random parameter model where variation in technology parameters accounts for heterogeneity (see, e.g. Kalirajan and Obwona, 1994; Tsionas, 2002; Huang, 2004). Often these random parameter models are formulated within the Bayesian framework with certain prior distributions for the parameters. Greene also discusses so-called latent class models or alternatively named 'finite mixture models' (see, e.g. Beard *et al*., 1991; Gropper *et al*., 1999; Caudill, 2003). These models assume that producers belong to different groups/classes for which different production functions with different technology parameters are estimated. Often the probability of class membership parametrized as a function of some firm-/environment-specific variables. Greene aptly points out that these latent class models can be seen as discrete versions of the random parameter models above. The main limitation of these models is that the number of classes has to be known beforehand. On the DEA side, group-specific frontiers are estimated in frontier separation (Charnes *et al*., 1981) and meta-frontier approaches (Battese *et al*., 2004; O'Donnell *et al*., 2008). These models however require that we beforehand know to which group each producer belongs. In the latent class models, the class membership is unknown *a priori*.

Instead of estimating separate technologies based on some categorization on environment or producer characteristics, we can consider that production technology differs at different levels of efficiency. It might be reasonable to expect that technology closer to the frontier differs from technology far from it. Quantile regression has been utilized for this purpose (Bernini *et al*., 2004; Liu *et al*., 2008; for a general treatment of quantile regression see Koenker and Hallock, 2001). Different quantiles are estimated such that one of the upper quantiles represents the efficient frontier. The problem is that the choice of appropriate quantile to represent the frontier is relatively arbitrary. Finally, the approach by Li *et al*. (2002) models technology heterogeneity by defining the production function parameters as functions of some environmental variables. This is yet another placement possibility for the z-variables.

In general, the aim of including technology heterogeneity in a stochastic frontier context can be summarized by the statement of Tsionas (2002): '*. . . free the frontier model from the restrictive assumption that all firms must share exactly the same technological possibilities*'. This would leave us to study the 'true' inefficiency that is remaining after accounting for technological differences. Moreover, since technological differences are arguably mainly due to differences in operating environment, it seems that the random parameter models could provide information on how technological choices respond to the changes in environment. However, if the aim is only to allow flexibility in the production function, we may want to resort to DEA style non-parametric methods, which inherently are very flexible in terms of the production technology.

The issue of heterogeneity is of course not new. Already Hall and Winsten (1959) considered possible problems in efficiency estimation when producers operating in different environments were compared. According to them, difficulties arise since each environment sets a different range of choices for managers. Interestingly, the current definition of z-variables can be dated back to their paper as they point out the difference between production and environment variables as: '*The lines between different classes of comparisons are drawn by those changes which do not count as changes of technique*, *but which do influence output*'. They also suggest that some 'allowances' should be made according to how difficult a certain task is to achieve in certain operating environment. This is basically the core reason why z-variables are used. For example, in comparing the cost efficiency of electricity distribution firms, it is important to take into account the operating environment. Companies in urban areas have to utilize more

expensive underground cabling instead of cheaper overhead cabling that can be used in more rural areas. Direct comparison of cost efficiency could then be unfair if the operating environment is left unaccounted. For example Kuosmanen (2012) and Kuosmanen *et al*. (2013) use the proportion of underground cabling from total cabling as a variable characterizing the operation environment.

The discussion above characterizes the multitude of ways by which we may account for the heterogeneity of producers or their environment. Many of them reach far beyond the concept of heteroscedasticity. However, whether heterogeneity is due to producer characteristics or environment characteristics can be seen only as secondary in importance. The crucial issue is to acknowledge that some form of heterogeneity is almost always present and neglecting it will be likely to lead to haphazard results and interpretations. Of course in practice, we need to specify what type of heterogeneity we are looking for, since from an estimation point of view it seems almost impossible or at least impractical to include all types of heterogeneity. Thus, it is worth emphasizing that none of the above ways to include heterogeneity is more correct than any other. The application at hand dictates what type of heterogeneity we ought to model and how to model it. Considering the modelling of risk, models of heteroscedasticity seem the most obvious choice.

6. Conclusions

This paper has examined two predominantly separate fields of research. Production risk literature has attempted to understand the distribution of output and its related risk considerations. Specifically, the field has contributed much effort to model how output variation could be better controlled through input use. Frontier literature has concentrated on estimating maximum obtainable outputs and identifying departures from this output due to inefficiency. At first these aims seem rather distinctive. However, with a systemic coverage of the relevant literature, this review has built a more coherent picture of the connections between these fields. The concept of heteroscedasticity has been utilized to bring these two veins of literature on par with each other. Many of the empirical methods in these fields can be connected via heteroscedasticity. Estimation methods that attempt to jointly estimate inefficiency and production risk have also been covered. In many instances, this joint estimation seems necessary as it is likely that neglecting the other factor might bias our estimation of the other. Unfortunately, conceptually it is difficult to identify between production risk, inefficiency and general producer heterogeneity. Empirically two alternatives exist. We can increase model complexity and try to estimate everything at once. Especially adding general producer heterogeneity and risk preferences may inhibit us from using simple models. Alternatively, we can take some traditional models and interpret their results in terms of risk and inefficiency. Many heteroscedastic SFA models covered in this review would fall in this category. Both approaches however aim for synthesis, either through novelty in estimation or in interpretation.

Although giving exact methodological prescriptions for future research is not the task of this review, one general suggestion is made. Considering heteroscedasticity, it is necessary to define whether we consider it only as an econometric problem or a concept with some economic meaning. In the former case, a correction of it or models robust to it are appropriate ways to proceed. However, if we aim to model the economic meaning of heteroscedasticity in a production economics context, we should probably look towards models of risk. On the other hand, if we are examining risk, it is the models of heteroscedasticity that we should first look towards. Furthermore, the analysis of risk in a frontier context could be extended with the concepts such as heteroskewness (Antle, 1983b; Bera and Sharma, 1999) and heterokurtosis since both skewness and kurtosis seem relevant for risk considerations.

In summary, this review reveals a clear connection between the production risk literature and the frontier literature. The fact that this connection mainly rests upon a single concept at the moment is something that we view as a potential for future development. Establishing further connections through methodological advances beyond the ones presented here is still a research agenda that could

be extended. Since these issues still seem unsettled, it is sufficient to end this review with the still relevant question that Anderson (1974) presented: '*Can proper account of risk be taken in research and extension*'?

Notes

1. The term *frontier field* is used in this study to refer to a branch of productivity and efficiency literature, which is interested in estimating production frontiers and deviations from this frontier.
2. Some may differentiate between risk and uncertainty (Knight 1921), but others may not (Chavas 2004). In our discussion we use the terms interchangeably.
3. More generally, production risk can be defined as uncertainty over output or its price. We restrict ourselves to output uncertainty and do not consider output price uncertainty. See, for example, Kumbhakar (2002b) references therein for an output price uncertainty case.
4. The difference in these results may also be explained by the fact that Day mainly focused on cotton crops whereas Fuller investigated corn crop. For example, Pannell (1991) has pointed that the effects of pesticide use does not always need to be risk reducing if multiple sources of output/income uncertainty are considered. It is possible that nitrogen use interacts with the multiple sources of risk differently in the respective cases of cotton and corn.
5. We have mainly presented the formulas as they appear in the original articles and have altered the notation only when possible confusion may occur.
6. Antle also noted that many other stochastic production models, namely stochastic frontier models with a two-component error term, impose restrictions such that all elasticities (w.r.t. inputs) of moments beyond the first moment are directly proportional to the elasticity of the first moment.
7. See, for example, Pannell (1991) and more recently Just and Pope (2003) for a discussion about the role of risk preferences in explaining risk responses in agriculture.
8. More generally, Hanoch and Levy (1969) notice that the mean-variance criterion is a sufficient condition for the comparison of efficiency of two risky prospects if both prospects follow a two-parameter distribution. Efficiency in their terminology refers to the dominance of one certain distribution over another. However, mean-variance criterion is not a necessary condition for efficiency. See also Anderson (1974) who considers stochastic dominance concepts in comparing technologies in an agricultural context.
9. Dominguez-Molina *et al*. (2003) state the stochastic frontier model in terms of a skew-normal distribution of the composed error term.
10. Tveterås (1999) found that technical change has contributed positively to the output risk in the Norwegian salmon farming industry. This is somewhat surprising but from the point of technology adoption, it is a plausible result as technology adoption might (temporarily) increase variability. Tveterås shows that the effect on mean production has dominated the variance effect, thus, implying an improved technical efficiency over all other producers, regardless of risk preferences.
11. Of course, any model is only an approximation of the production process, as Jolly (1983) points out: '*Generally speaking no matter how many stochastic or dynamic bells and whistles are added to the optimization problem, it will likely remain a stylized and incomplete representation of the management process*'. However, we could argue that, if indeed, such a complete representation of the management process accounting for every manager and operating environment-specific factor could be constructed, econometrically no producer would seem inefficient.
12. Inputs and, thus, outputs are *state-allocable* in their model. The period 0 output is non-stochastic and the period 1 output is stochastic. The task for a producer is to allocate his production in an optimal way given the probabilities of the states of nature.

References

Abramovitz, M. (1956) Resource and output trends in the United States since 1870. *American Economic Review* 46(2): 5–23.

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

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

Amemiya, T. (1977) A note on a heteroscedastic model. *Journal of Econometrics* 6: 365–370.

Amsler C., Lee, Y.H. and Schmidt, P. (2009) A survey of stochastic frontier models and likely future developments. *Seoul Journal of Economics* 22(1): 5–27.

Anderson, J. (1973) Sparse data, climatic variability, and yield uncertainty in response analysis. *American Journal of Agricultural Economics* 55(1): 77–82.

Anderson, J. (1974) Risk efficiency in the interpretation of agricultural production research. *Review of Marketing and Agricultural Economics* 42(3): 131–184.

Antle, J.M. (1983a) Incorporating risk in production analysis. *American Journal of Agricultural Economics* 65(5): 1099–1106.

Antle, J.M. (1983b) Testing the stochastic structure of production: a flexible moment-based approach. *Journal of Business and Economic Statistics* 1(3): 192–201.

Antle, J.M. (2010a) Asymmetry, partial moments, and production risk. *American Journal of Agricultural Economics* 92(5): 1294–1309.

Antle, J.M. (2010b) Do economic variables follow scale or location-scale distributions? *American Journal of Agricultural Economics* 92(1): 196–204.

Antle, J. and Goodger, W. (1984) Measuring stochastic technology – the case of Tulare milk-production. *American Journal of Agricultural Economics* 66(3): 342–350.

Antle, J.M. and Crissman, C.C. (1990) Risk, efficiency, and the adoption of modern crop varieties: evidence from the Philippines. *Economic Development and Cultural Change* 38(3): 517–537.

Asche, F. and Tveteras, R. (1999) Modeling production risk with a two-step procedure. *Journal of Agricultural and Resource Economics* 24(2): 424–439.

Battese, G.E. (1992) Frontier production functions and technical efficiency: a survey of empirical applications in agricultural economics. *Agricultural Economics* 7(3–4): 185–208.

Battese, G.E. and Coelli, T.J. (1995) A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics* 20(2): 325–332.

Battese, G.E., Rambaldi, A.N. and Wan, G.H. (1997) A stochastic frontier production function with flexible risk properties. *Journal of Productivity Analysis* 8(3): 269–280.

Battese, G.E., Rao, D. and O'Donnell, C. (2004) A metafrontier production function for estimation of technical efficiencies and technology gaps for firms operating under different technologies. *Journal of Productivity Analysis* 21(1): 91–103.

Beard, T.R., Caudill, S.B. and Gropper, D.M. (1991) Finite mixture estimation of multiproduct cost functions. *The Review of Economics and Statistics* 73(4): 654–664.

Bera, A.K. and Sharma, S.C. (1999) Estimating production uncertainty in stochastic frontier production function models. *Journal of Productivity Analysis* 12(3): 187–210.

Bernini, C., Freo, M. and Gardini, A. (2004) Quantile estimation of frontier production function. *Empirical Economics* 29: 373–381.

Bravo-Ureta, B.E., Solís, D., Moreira López, V.H., Maripani, J.F., Thiam, A. and Rivas, T. (2007) Technical efficiency in farming: a meta-regression analysis. *Journal of Productivity Analysis* 27(1): 57–72.

Caudill, S.B. (2003) Estimating a mixture of stochastic frontier regression models via the EM algorithm: a multiproduct cost function application. *Empirical Economics* 28(3): 581–598.

Caudill, S. and Ford, J. (1993) Biases in frontier estimation due to heteroscedasticity. *Economics Letters* 41(1): 17–20.

Caudill, S., Ford, J. and Gropper, D. (1995) Frontier estimation and firm-specific inefficiency measures in the presence of heteroscedasticity. *Journal of Business and Economic Statistics* 13(1): 105–111.

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

Charnes, A., Cooper, W.W. and Rhodes, E. (1981) Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through. *Management Science* 27(6): 668–697.
Chavas, J.P. (2004). *Risk Analysis in Theory and Practice*. San Diego: Elsevier Academic Press.
Daraio, C. and Simar, L. (2005) Introducing environmental variables in nonparametric frontier models: a probabilistic approach. *Journal of Productivity Analysis* 24(1): 93–121.
Day, R. (1965) Probability-distributions of field crop yields. *Journal of Farm Economics* 47(3): 713–741.
Debreu, G. (1951) The coefficient of resource utilization. *Econometrica* 19(3): 273–292.
Domínguez-Molina, J.A., Gonzáles-Farías, G. and Ramos-Quiroga, R. (2003) Skew-normality in stochastic frontier analysis. *Comunicacion Tecnica* I-03–18:1–13. Available at http://www.cimat.mx/reportes/enlinea/I-03--18.pdf (last accessed 21 September 2012).
Du, X., Hennessy, D.A. and Yu, C.L. (2012) Testing Day's conjecture that more nitrogen decreases crop yield skewness. *American Journal of Agricultural Economics* 94(1): 225–237.
Farrell, M.J. (1957) The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)* 120(3): 253–290.
Fried, H.O., Lovell, C.A.K. and Schmidt, S.S., eds. (2008) *The Measurement of Productive Efficiency and Productivity Growth*. New York: Oxford University Press Inc.
Fuller, W. (1965) Stochastic fertilizer production-functions for continuous corn. *Journal of Farm Economics*, 47(1): 105–119.
Gallagher, P. (1987) U.S. soybean yields: estimation and forecasting with nonsymmetric disturbances. *American Journal of Agricultural Economics* 69(4): 796–803.
Ghosh, S., McGuckin, J.T. and Kumbhakar, S.C. (1994) Technical efficiency, risk attitude, and adoption of new technology: the case of the U.S. dairy industry. *Technological Forecasting and Social Change* 46(3): 269–278.
Giannakis, D., Jamasb, T. and Pollitt, M. (2005) Benchmarking and incentive regulation of quality of service: an application to the UK electricity distribution networks. *Energy Policy* 33(17): 2256–2271.
Greene, W. (2005) Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics* 126(2): 269–303.
Greene, W.H. (2004) Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organizations's panel data on national health care systems. *Health Economics* 13: 959–980.
Greene, W.H. (2008) The econometric approach to efficiency analysis. In H.O. Fried, C.A.K. Lovell and S.S. Schmidt (eds.), *The Measurement of Productive Efficiency and Productivity Growth* (pp. 92–250). New York: Oxford University Press Inc.
Griffiths, W.E. and Anderson, J.R. (1982) Using time-series and cross-section data to estimate a production function with positive and negative marginal risks. *Journal of the American Statistical Association* 77(379): 529–536.
Gropper, D.M., Caudill, S.B. and Beard, T.R. (1999) Estimating multiproduct cost functions over time using a mixture of normals. *Journal of Productivity Analysis* 11(3): 201–218.
Hadri, K. (1999) Estimation of a doubly heteroscedastic stochastic frontier cost function. *Journal of Business and Economic Statistics* 17(3): 359–363.
Hall, M. and Winsten, C. (1959) The ambiguous notion of efficiency. *Economic Journal* 69(273): 71–86.
Hanoch, G. and Levy, H. (1969) The efficiency analysis of choices involving risk. *The Review of Economic Studies* 36(3): 335–346.
Harvey, A.C. (1976) Estimating regression models with multiplicative heteroscedasticity. *Econometrica* 44(3): 461–465.
Huang, C.J. and Liu, J. (1994) Estimation of a non-neutral stochastic frontier production function. *Journal of Productivity Analysis* 5(2): 171–180.
Huang, H. (2004) Estimation of technical inefficiencies with heterogeneous technologies. *Journal of Productivity Analysis* 21(3): 277–296.
Huang, T. and Kao, T. (2006) Joint estimation of technical efficiency and production risk for multi-output banks under a panel data cost frontier model. *Journal of Productivity Analysis* 26(1): 87–102.
Jaenicke, E., Frechette, D. and Larson, J. (2003) Estimating production risk and inefficiency simultaneously: an application to cotton cropping systems. *Journal of Agricultural and Resource Economics* 28(3): 540–557.

Jolly, R.W. (1983) Risk management in agricultural production. *American Journal of Agricultural Economics* 65(5): 1107–1113.
Jondrow, J., Lovell, C., Materov, I. and Schmidt, P. (1982) On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics* 19(2–3): 233–238.
Just, R.E. (1974) An investigation of the importance of risk in farmers' decisions. *American Journal of Agricultural Economics* 56(1): 14–25.
Just, R.E. and Just, D.R. (2011) Global identification of risk preferences with revealed preference data. *Journal of Econometrics* 162: 6–17.
Just, R.E. and Pope, R.D. (1978) Stochastic specification of production functions and economic implications. *Journal of Econometrics* 7(1): 67–86.
Just, R.E. and Pope, R.D. (1979) Production function estimation and related risk considerations. *American Journal of Agricultural Economics* 61(2): 276–284.
Just, R.E. and Pope, R.D. (2001) The agricultural producer: theory and statistical measurement. In B.L. Gardner and G.C. Rausser (eds.), *Handbook in Agricultural Economics* (Vol. 1, Part 1, pp. 629–741). Amsterdam: Elsevier. Available at http://www.sciencedirect.com/science/handbooks/15740072 (last accessed 7 September 2011).
Just, R.E. and Pope, R.D. (2003) Agricultural risk analysis: adequacy of models, data, and issues. *American Journal of Agricultural Economics* 85(5): 1249–1256.
Just, R.E. and Weninger, Q. (1999) Are crop yields normally distributed? *American Journal of Agricultural Economics* 81(2): 287–304.
Kalirajan, K. and Obwona, M. (1994) Frontier production function – the stochastic coefficients approach. *Oxford Bulletin of Economics and Statistics* 56(1): 87–96.
Kendall, M. and Stuart, A. (1977) *The Advanced Theory of Statistics*, Vol. 1, 4th edn, New York: Macmillan.
Knight, F. (1921) *Risk, Uncertainty and Profit*. Boston: Houghton Mifflin.
Koenker, R. and Hallock, K.F. (2001) Quantile Regression. *Journal of Economic Perspectives* 17(4): 143–156.
Koopmans, T.C. (1951) An analysis of production as an efficient combination of activities. In T.C. Koopmans (ed.) *Activity Analysis of Production and Allocation*. Cowles Commission for Research Monograph No. 13. New York: John Wiley and Sons.
Kopp, R.J. and Mullahy, J. (1990) Moment-based estimation and testing of stochastic frontier models. *Journal of Econometrics* 46(1–2): 165–183.
Koundouri, P. and Nauges, C. (2005) On production function estimation with selectivity and risk considerations. *Journal of Agricultural and Resource Economics* 30(3): 597–608.
Kumbhakar, S. and Tveteras, R. (2003) Risk preferences, production risk and firm heterogeneity. *Scandinavian Journal of Economics* 105(2): 275–293.
Kumbhakar, S.C. (1993) Production risk, technical efficiency, and panel data. *Economics Letters* 41: 11–16.
Kumbhakar, S.C. (2002a) Specification and estimation of production risk, risk preferences and technical efficiency. *American Journal of Agricultural Economics* 84(1): 8–22.
Kumbhakar, S.C. (2002b) Risk preference and productivity measurement under output price uncertainty. *Empirical Economics* 27(3): 461–472.
Kumbhakar, S.C., and Lovell, C.A.K. (2000) *Stochastic Frontier Analysis*. New York, USA: Cambridge University Press.
Kumbhakar, S.C. and Tsionas, E.G. (2010) Estimation of production risk and risk preference function: A nonparametric approach. *Annals of Operations Research* 176: 369–378.
Kumbhakar, S.C., Ghosh, S., and McGuckin, J. (1991) A generalized production frontier approach for estimating determinants of inefficiency in U.S. dairy farms. *Journal of Business and Economic Statistics* 9(3): 279–286.
Kumbhakar, S.C., Park, B.U., Simar, L., and Tsionas, E.G. (2007) Nonparametric stochastic frontiers: A local maximum likelihood approach. *Journal of Econometrics* 137(1): 1–27.
Kuosmanen, T. (2012). Stochastic semi-nonparametric frontier estimation of electricity distribution networks: Application of the StoNED method in the Finnish regulatory model. *Energy Economics* 34: 2189–2199.
Kuosmanen, T. and Fosgerau, M. (2009) Neoclassical versus frontier production models? Testing for the skewness of regression residuals. *Scandinavian Journal of Economics* 111(2): 351–367.
Kuosmanen, T., Saastamoinen, A. and Sipiläinen, T. (2013). What is the best practice for benchmark regulation of electricity distribution? Comparison of DEA, SFA and StoNED methods. *Energy Policy* 61: 740–750.

Lence, S.H. (2009) Joint estimation of risk preferences and technology: flexible utility or futility? *American Journal of Agricultural Economics* 91(3): 581–598.
Li, Q., Huang, C., Li, D. and Fu, T. (2002) Semiparametric smooth coefficient models. *Journal of Business and Economic Statistics* 20(3): 412–422.
Liu, C., Laporte, A., and Ferguson, B.S. (2008) The quantile regression approach to efficiency measurement: insights from Monte Carlo simulations. *Health Economics* 17(9): 1073–1087.
Love, H. and Buccola, S. (1991) Joint risk preference-technology estimation with a primal system. *American Journal of Agricultural Economics* 73(3): 765–774.
Meeusen, W. and van den Broeck, J. (1977) Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review* 18(2): 435–445.
Moschini, G. and Hennessy. D.A. (2001) Uncertainty, risk aversion, and risk management for agricultural producers. In B.L. Gardner and G.C. Rausser (eds.) *Handbook in Agricultural Economics* (Vol. 1, Part 1, pp. 88–153). Amsterdam: Elsevier. Available at http://www.sciencedirect.com/science/handbooks/15740072 (last accessed 7 September 2011).
Murillo-Zamorano, L. (2004) Economic efficiency and frontier techniques. *Journal of Economic Surveys* 18(1): 33–77.
O'Donnell, C. and Griffiths, W. (2006) Estimating state-contingent production frontiers. *American Journal of Agricultural Economics* 88(1): 249–266.
O'Donnell, C., Rao, D. and Battese, G.E. (2008) Metafrontier frameworks for the study of firm-level efficiencies and technology ratios. *Empirical Economics* 34(2): 231–255.
O'Donnell, C.J., Chambers, R.G. and Quiggin, J. (2010) Efficiency analysis in the presence of uncertainty. *Journal of Productivity Analysis* 33(1): 1–17.
Pannell, D. (1991) Pests and pesticides, risk and risk-aversion. *Agricultural Economics* 5(4): 361–383.
Pope, R. and Chavas, J. (1994) Cost-functions under production uncertainty. *American Journal of Agricultural Economics* 76(2): 196–204.
Pope, R. and Just, R.E. (1998) Cost function estimation under risk aversion. *American Journal of Agricultural Economics* 80(2): 296–302.
Ramaswami, B. (1992). Production risk and optimal input decisions. *American Journal of Agricultural Economics* 74(4): 860–869.
Reifschneider, D. and Stevenson, R. (1991) Systematic departures from the frontier – a framework for the analysis of firm inefficiency. *International Economic Review* 32(3): 715–723.
Saha, A., Havenner, A. and Talpaz, H. (1997) Stochastic production function estimation: small sample properties of ML versus FGLS. *Applied Economics* 29(4): 459–469.
Schmidt, P. (2010) One-step and two-step estimation in SFA models. *Journal of Productivity Analysis* 36(2): 201–203.
Schmidt, P. and Sickles, R. (1984) Production frontier and panel data. *Journal of Business & Economic Statistics* 2(4): 367–374.
Simar, L., Lovell ,C.A.K. and Vanden Eeckaut, P. (1994) Stochastic frontiers incorporating exogenous influences on efficiency. Discussion Papers No. 9403, Institut de Statistique, University Catholique de Louvain.
Stigler, G.J. (1976) The xistence of X-inefficiency. *American Economic Review* 66(1): 213–216.
Traxler, G., Falck-Zepeda, J., J.I. Ortiz-Monasterio, R. and Sayre, K. (1995) Production risk and the evolution of varietal technology. *American Journal of Agricultural Economics* 77(1): 1–7.
Tsionas, E. (2002) Stochastic frontier models with random coefficients. *Journal of Applied Econometrics* 17(2): 127–147.
Tveterås, R. (1999) Production risk and productivity growth: some findings for Norwegian salmon aquaculture. *Journal of Productivity Analysis* 12(2): 161–179.
Varian, H. (1992). *Microeconomic Analysis*. New York: W.W. Norton & Company Inc.
Wang, H. (2002) Heteroscedasticity and non-monotonic efficiency effects of a stochastic frontier model. *Journal of Productivity Analysis* 18(3): 241–253.
Wang, H. and Schmidt, P. (2002) One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. *Journal of Productivity Analysis* 18(2): 129–144.

Article 4

This is the authors accepted manuscript of an article published as the version of record in Applied Economics May 2014

Antti Saastamoinen; Timo Kuosmanen. Is Corruption Grease, Grit, or a Gamble? Corruption Increases Variance of Productivity Across Countries. Applied Economics, 2014, vol. 46, Issue 23, pp. 2833-2849. http://www.tandfonline.com/doi/full/10.1080/00036846.2014.914149

Is Corruption Grease, Grit, or a Gamble?
Corruption Increases Variance of Productivity Across Countries

Antti Saastamoinen, Timo Kuosmanen

ABSTRACT

The effect of corruption on economic growth has attracted interest in empirical development economics. The conventional view of corruption as impediment for growth has been challenged by the "grease-on-the-wheels" hypothesis. We take a new perspective on the issue and suggest corruption as macro risk, referred to as a "gamble" hypothesis. Using cross-country data and two alternative indicators of corruption, we find corruption to be a significant driver of heteroskedasticity in total productivity. This supports the new gamble hypothesis. We also note some misleading interpretations in the previously published frontier applications. To avoid these shortcomings, we apply a flexible semi-nonparametric estimator.

KEYWORDS: productivity; corruption; heteroskedasticity; economic growth; semi-parametric regression

JEL codes: O47, O17, C14

I. INTRODUCTION

While the importance of institutional factors on macroeconomic performance is widely recognised in economics (Hall & Jones, 1999), the effects of corruption on the growth and productivity remain a subject of debate. There is vast empirical evidence that corruption is detrimental to investments and to the economic performance and development (Mauro, 1995, 1998; Mo, 2001; Lambsdorff, 2003, 2005; Assane & Grammy, 2003; Méon & Sekkat, 2005; Everhart et al., 2009; see also Haggard & Tiede, 2010 and the references therein). On the other hand, the possibility of "efficient corruption" has already been suggested in the works by Leff (1964), Leys (1965), and Huntington (1968), among others. Subsequently, the positive effect of corruption on economic performance has been commonly labeled as the "*grease in the wheels"* hypothesis (see for example, Méon & Weill, 2010). Although the vast majority of empirical research refutes the grease hypothesis, it has stayed as a topic of discussion among researchers due to the casual real-life observations of this phenomenon. China's impressive economic growth provides an obvious example: China scores 3.6 out of 10 in the Corruption Perception Index of 2011 by Transparency International (score 10 indicates little corruption). Other examples of the so-called Asian paradox are India, South Korea, and Indonesia, which have experienced high growth during arguably a corrupt regime or institutional setting (Khan, 1996; Bardhan, 1997; Heston & Kumar, 2007; Vial & Hanoteau, 2010). In general the more recent literature suggests that the effects of corruption on growth are very much dependent from the surrounding institutional setting (see for example, Bardhan, 2006; Méndez & Sepúlveda, 2006; Méon & Weill, 2010).

The intuition behind the grease hypothesis is that bribery or corruption might act as a lubricant of otherwise rigid bureaucratic system (Beenstock, 1979; Lui, 1985). It has been also suggested that some centralised institutional frameworks, even though highly prone to corruption, can be good for economic performance if the institutions are well organised and harnessed to the clear objective of growth (Ehrlich & Lui, 1999). Consequently some form of centralised corruption regime might have some merits over de-centralised corruption (Shleifer &Vishny, 1993; Blackburn & Forgues-Puccio, 2009). Furthermore Nye (1967) was one of the first to suggest that corruption might act as a temporary fix to achieve certain development goals.

In this paper we argue that corruption is not simply grease or grit in the wheels, but perhaps more importantly, we consider corruption as a risk factor for economic performance (productivity) in the development process. We refer to the proposed risk interpretation as the "*gamble hypothesis*" of corruption. The risk interpretation warrants us to study the effects of corruption on the variance of productivity, in the spirit of the seminal work on production risk by Just and Pope

(1978). To this end, we draw a clear distinction between the effects of corruption on the level (that is, the grease or the grit) and the variance of productivity (the new gamble hypothesis). We introduce the grease, grit, and gamble hypotheses formally, and test them on empirical data.

To test the gamble hypothesis, we can utilise the established econometric tools for modeling heteroskedasticity. In the terminology of econometrics, the gamble hypothesis implies that the distribution of productivity is heteroskedastic with respect to corruption. Although heteroskedasticity is a major issue in econometrics, the empirical literature on corruption and growth has thus far largely ignored it. To test for the gamble hypothesis we take the heteroskedasticity effects of corruption systematically into account. Our empirical results indicate that corruption has a significant positive correlation with the variance of productivity across countries.

Our secondary objective is to clarify an econometric issue that has apparently caused some confusion in the empirical literature on corruption. More specifically, we show in Section 3 that in stochastic frontier models applied in this literature, the effects of corruption on the level and the variance of productivity are indistinguishable. Thus, the heteroskedasticity effects may have been wrongly interpreted as support to the grit or the grease hypotheses. Therefore, while the significant heteroskedasticity effect of corruption is an interesting empirical finding as such, it can also help us to understand and reconcile seemingly contradictory empirical evidence in favor and against the grease hypothesis.

As our third goal, we examine the sensitivity of the estimated effects of corruption on the restrictive parametric specification of the aggregate production functions. We apply a new semi-nonparametric frontier estimator developed by Kuosmanen and Kortelainen (2012) and Johnson and Kuosmanen (2011, 2012), which allows us to estimate the effects of corruption on productivity without imposing arbitrary assumptions on the functional form of the production function or the probability distribution of the composite disturbance term. This approach allows the output elasticities of production factors differ across countries, in the spirit of Durlauf and Johnson (1995). Further, this estimator allows for stochastic noise, and thus circumvents the usual shortcoming of the conventional deterministic non-parametric frontier estimators (Kumar & Russell, 2002; footnote 7).

The rest of the paper is organised as follows. Section 2 briefly reviews the theoretical literature on which the gamble hypothesis is based on, and states the hypotheses to be tested. Section 3 presents the alternative estimation methods considered in this paper. Section 4 describes the data and its sources. Section 5 presents the empirical results. Section 6 presents our concluding remarks and suggests avenues for further research. Finally, we provide the program code and the data used

in our empirical estimations, and some further results and graphical illustrations in the supplementary material that is available online.

II. MODELING FRAMEWORK AND THEORETICALBACKGROUND

Consider the standard macroeconomic model of aggregate production, which can be stated as

$$y = \phi(\mathbf{x})\exp(\varepsilon) \tag{1}$$

where y is the total economic output of the country (GDP), $\mathbf{x}$ is the vector of production factors (capital, labor), and $\phi(\mathbf{x})$ is the aggregate production function. Random variable ε represents productivity shocks due to technological change and inefficiency. In addition, the random variable ε captures effects of any omitted or unobserved variables, which do not appear as inputs to the production function. In this paper we use the term *contextual variable* to refer to institutional and geographic factors that are not production factors as such, but which nevertheless influence the expected value and variance of the random variable ε.[1] Denoting the vector of contextual variables by $\mathbf{z}$, we assume that $E(\varepsilon) = \boldsymbol{\delta}'\mathbf{z}$ and $Var(\varepsilon) = \boldsymbol{\gamma}'\mathbf{z}$. In words, we assume that both the expected value and the variance of the productivity shock depends on the contextual variables $\mathbf{z}$. Coefficients $\boldsymbol{\delta}$ represent the marginal effects of contextual variables on the level of productivity, whereas coefficients $\boldsymbol{\gamma}$ represent the marginal effects on the variance.

In this paper we are mainly interested in the effects of corruption, which we model as one of the contextual variables. Specifically, the first element of vector $\mathbf{z}$, denoted as z_1, is reserved for an index of corruption. We assume that an increase in z_1 implies decrease in corruption, analogous to the World Bank's Governance Indicators (WGI) and the Corruption Perception Index (CPI) by the Transparency International. Our rationale for treating corruption as a contextual variable is the following. Although corruption obviously influences the economic output, it is difficult to control the corruption by the government policy. Further, the effects of corruption on output are far from deterministic. The uncertain influence of corruption is explicitly recognised in our probabilistic model where corruption affects output indirectly through the mean and the variance of the random variable ε.

Taking the logarithms of both sides of (1), we can rearrange the terms to obtain

$$\ln\left(\frac{y}{\phi(\mathbf{x})}\right) = \varepsilon \tag{2}$$

Interpreting $\phi(\mathbf{x})$ as an input index (aggregator of production factors $\mathbf{x}$), we have a compelling interpretation of the left hand side of Equation (2) as the logarithm of total factor productivity (TFP). Therefore, we find it appropriate to interpret the random variable ε as a TFP shock, which may be due to technological innovations, inefficiency, or institutional factors, among other reasons.[2] The distinction between the TFP growth rates (convergence) and the levels of TFP is worth noting. As institutional change is very slow, Hall and Jones (1999) consider the level of TFP to be more relevant in the present context. The TFP levels represent the economic performance of the country in the long-run, whereas productivity change is only 'transitory' in its nature. On the other hand, Cherchye and Moesen (2003) and Salinas-Jiménez and Salinas-Jiménez (2010) find that that effects of institutional factors on the level of TFP and its growth rate are very similar. In this study we focus on the TFP levels, noting that our approach could be applied to the study of growth rates in a straightforward manner.

We are now equipped to state the three empirical hypotheses to be tested, defined as follows:

Grease hypothesis: $\dfrac{\partial E(\varepsilon)}{\partial z_1} = \delta_1 < 0$

Grit hypothesis: $\dfrac{\partial E(\varepsilon)}{\partial z_1} = \delta_1 > 0$

Gamble hypothesis: $\dfrac{\partial Var(\varepsilon)}{\partial z_1} = \gamma_1 < 0$

Since ε is a random variable, the above hypotheses are stated in terms of the expected value and variance.

The grease hypothesis suggests that an increase in the index z_1 (reduction in corruption) decreases the expected value of logarithm of TFP. The grit hypothesis implies the opposite effect, that reduction in corruption increases the expected TFP. In this interpretation, the grease and grit hypotheses are mutually exclusive statements regarding the expected value. The accumulated empirical evidence mainly supports the grit hypothesis (for example, Adkins, Moomaw & Savvides, 2002; Cherchye & Moesen, 2003; Salinas-Jiménez & Salinas-Jiménez, 2007, 2010; Hauner & Kyobe, 2010). However, for example Méon & Weill, (2010) and Rock & Bonnett (2004) do find some support for the grease hypothesis. Méon & Weill, (2010) nevertheless differentiate with two forms of the grease hypothesis, namely strong and weak form. The strong form says that at certain low levels of institutional quality, corruption can have positive effect on productivity. The weak form however states that corruption is

only less detrimental at certain low levels of institutional quality, but overall the effect on productivity is negative for all countries.

The gamble hypothesis introduced in this paper considers the effect of corruption on the variance of TFP, stating that corruption increases the variance, which can be further interpreted as increase in risk. As an empirical motivation, we refer to Figure 1 in Wyatt (2003), which presents a scatter-plot of a corruption index and the logarithm of labor productivity. This figure aptly illustrates how the variance of productivity across countries is associated with the level of corruption. In a similar vein, Cavalcanti and Álvaro (2005) observe that: "… *the distribution of output per worker tends to become less disperse as countries improve their institutional framework*." To our knowledge, however, the gamble hypothesis or the heteroskedasticity effects of corruption have not been systematically examined or tested statistically before.

The theoretical backbone of the gamble hypothesis can be drawn from the literature of multiple growth regimes. Durlauf and Johnson (1995) empirically refute the standard linear growth model in favor of a multiple regime growth model. In their view, different growth regimes emerge as the initial level of economic development differs from one country to another. Thus, their findings conform to the multiple steady state growth models. The idea of multiple regimes has carried over to the empirical literature of the effects of institutional factors on growth (Méndez & Sepúlveda, 2006; Aidt, Dutta & Sena, 2008; Méon & Weill, 2010). Recently, Blackburn and Forgues-Puccio (2009) present a model where the effect of corruption varies between the different levels of coordination of corrupt practices among bureaucrats, conforming to the literature on centralised versus de-centralised corruption. This issue can also be related to corruption club literature where countries experience rather different equilibrium levels of corruption (Herzfeld & Weiss, 2007). This is often explained by the complex interplay between other (legal) institutions and corruption (Herzfeld & Weiss, 2003). In the context of our analysis, the gamble analysis means that within a certain corruption club (i.e. equal corruption level) the effects of corruption on growth can vary.

The marginal effects presented above follow from our assumptions regarding the expected value and the variance of ε. The connection to the theory of multiple regimes can be established as follows. Suppose there exists only a single growth regime that does not depend on the level of corruption. This would imply that all countries have the same expected value and variance of productivity regardless of the corruption level, so we must have $\partial E(\varepsilon)/\partial z_1 = 0$ and $\partial Var(\varepsilon)/\partial z_1 = 0$. Now, suppose there exists multiple regimes of growth and productivity, which depend on level of corruption. This would suggest that the variance of productivity is not

constant, but we must have $\partial Var(\varepsilon)/\partial z_1 \neq 0$. Our gamble hypothesis additionally states that the sign of the variance effect is negative, such that corruption increases the risk.

III. ECONOMETRIC METHODS

The purpose of this section is to introduce the econometric methods that we apply for estimating the effects of corruption on TFP, and to clarify some issues that are well understood in econometrics, but have apparently caused some confusion in the empirical literature on corruption.

Our TFP measure includes the GDP as the output (*y*) and the capital stock, human capital, and labor inputs as the factors of production (**x**). The first approach to estimating the production function is the standard linear regression, which is used as a benchmark. It serves to study the effects of functional form on the significance of the estimates. The second method is the *Stochastic Frontier Analysis* (SFA hereafter) which is commonly applied in the empirical literature of frontier estimation.[3] As the third approach we consider the one-stage StoNED-estimator recently developed by Kuosmanen and Kortelainen (2012) and Johnson and Kuosmanen (2011, 2012).[4] Since SFA and particularly StoNED are not standard techniques, it is worth to briefly review these methods to enable readers to understand how the empirical results to be presented in Section 5 are obtained.

OLS regression

Given the general production function model in (1), we define our baseline model as the Cobb-Douglas production function with the linear effects of the contextual variables. For the sample of n countries indexed as $i = 1,...,n$, the baseline model is

$$\ln y_i = \alpha + \boldsymbol{\beta}'(\ln \mathbf{x}_i) + \boldsymbol{\delta}'\mathbf{z}_i + \tilde{\varepsilon}_i \tag{3}$$

where vectors are defined as in Section 2. The notation $\ln \mathbf{x}_i$ is used to indicate that the natural logarithm has been applied to each element of this vector. Notice that the disturbance term $\tilde{\varepsilon}_i$ is not the same as the random variable ε. Recall that we assumed $E(\varepsilon) = \boldsymbol{\delta}'\mathbf{z} \neq 0$. However, since the expected value of the productivity shock is explicitly controlled for in the regression equation, we can state the disturbance term as $\tilde{\varepsilon} = \varepsilon - \mathbf{z}'\boldsymbol{\delta}$ to obtain $E(\tilde{\varepsilon}) = E(\varepsilon) - \boldsymbol{\delta}'\mathbf{z} = \boldsymbol{\delta}'\mathbf{z} - \boldsymbol{\delta}'\mathbf{z} = 0$. We assume that the disturbances $\tilde{\varepsilon}_i$ are statistically independent of the inputs **x** and the contextual variables **z**, but we do allow $\tilde{\varepsilon}_i$ to be heteroskedastic. That is, the variances of the disturbance term can differ across countries.

We estimate the baseline model (1) by OLS. It is well known that the OLS estimator is unbiased and consistent under heteroskedasticity. A more efficient generalised least squares estimator could be used, but this requires that some specific model of heteroskedasticity is assumed. The main problem of the conventional OLS estimator is that the standard errors are invalid. To avoid this problem, we resort to White's (1980) heteroskedasticity robust standard errors in our statistical inferences based on OLS.

Stochastic Frontier Analysis

We next consider the stochastic frontier model by Kumbhakar, Ghosh and McGuckin (1991, KGM hereafter), where contextual variables such as corruption can be used as explanatory factors for productive inefficiency. In the present context, the KGM model has been used for example by Méon and Weill (2005) to study the effect of corruption on productivity. The KGM model has been adapted to the panel data setting by Battese and Coelli (1995). This panel data variant has been applied in several empirical studies on corruption and governance (Adkins et al., 2002; Jayasuriya & Wodon, 2005; Weill, 2008; and Méon & Weill, 2010).

In the stochastic frontier model, we use the Cobb-Douglas production function similar to the linear regression model in (1). However, the definition of the disturbance term differs from OLS. In the SFA model, the disturbance term consists of two components according to $\tilde{\varepsilon}_i = v_i - u_i$, where v_i is a random, normally distributed noise term with the zero mean and the variance σ_v^2, and $u_i > 0$ is an asymmetric inefficiency term. In the KGM model, the inefficiency term u_i depends on the contextual variables according to the following inefficiency model: $u_i = \boldsymbol{\theta}'\mathbf{z}_i + w_i$. We use the vector $\boldsymbol{\theta}$ to represent the indirect effects of the contextual variables through the inefficiency term u_i, to distinguish them from the direct effect modeled by the coefficients $\boldsymbol{\delta}$ as in Equation (3). The random variable w_i has a truncated normal distribution with the zero mean and the variance σ_w^2. The point of truncation is $-\boldsymbol{\theta}'\mathbf{z}_i$ such that $w_i \geq -\boldsymbol{\theta}'\mathbf{z}_i$. Thus, inefficiency u_i is distributed as a non-negative truncation of $N(\boldsymbol{\theta}'\mathbf{z}_i, \sigma_w^2)$. Consequently, the KGM specification of the SFA model can be stated as

$$\ln y_i = \alpha + \boldsymbol{\beta}'(\ln \mathbf{x}_i) + v_i - (\boldsymbol{\theta}'\mathbf{z}_i + w_i). \quad (4)$$

In the KGM model, the parameters of production frontier and the parameters of the inefficiency model are simultaneously estimated using maximum likelihood.

It might be tempting to interpret the parameters $\boldsymbol{\theta}$ of the SFA model as the marginal effects on the expected level of the productivity shock, similar to coefficients $\boldsymbol{\delta}$ in the regression Equation (3). However, this interpretation is incorrect and misleading, because coefficients $\boldsymbol{\theta}$ also influence the variance of the truncated inefficiency term u_i. This can be seen from the first two moments of the truncated normal distribution of u_i that can be stated as (see for example, Greene, 2003, p. 759):

$$\begin{aligned} E[u \mid truncation] &= \boldsymbol{\theta}'\mathbf{z}_i + \sigma_w \lambda(\omega) \\ Var[u \mid truncation] &= \sigma_w^2[1 - \upsilon(\omega)] \end{aligned}, \qquad (5)$$

where

$$\begin{aligned} \omega &= (-\boldsymbol{\theta}'\mathbf{z}_i) / \sigma_w \\ \lambda(\omega) &= \phi(\omega) / [1 - \Phi(\omega)] \\ \upsilon(\omega) &= \lambda(\omega)[\lambda(\omega) - \omega] \end{aligned}$$

In the set of Equations (5), $\phi(\omega)$ and $\Phi(\omega)$ are the standard normal density and cumulative distribution functions, respectively. The formal expression of the variance of the truncated random variable u_i reveals that the contextual variables $\mathbf{z}$ influences both the expected value and the variance of the inefficiency term. Therefore, the interpretation of the coefficients $\boldsymbol{\theta}$ is somewhat ambiguous: it is possible that they capture the effect of contextual variables on the expected level or the variance of productivity, or possibly both.[5] Therefore, we cannot use this model for testing the grease, grit, or the gamble hypotheses stated in Section 2.

Unfortunately, the truncation of the inefficiency term has not been duly recognised in the empirical literature on corruption and productivity. For example, Méon and Weill (2010) apply the panel data variant of the KGM model to test for the grit and grease hypotheses, but they overlook the truncation. Following Méon and Sekkat (2005), they define the effect of corruption being a function of other governance indicators by including an interaction between corruption index and other governance indicators. Interaction term indeed allows to test whether the effect of corruption depends from the other institutional setting. However the marginal effect of corruption on output becomes even more tedious to assess because one should also take into account the effect of the interaction terms to the truncation. Hence, it is possible that reported empirical results obtained using the KGM model or its panel data variants have been misinterpreted as support for the grit or the grease hypothesis, while in fact the coefficients $\boldsymbol{\theta}$ capture the heteroskedasticity in empirical production data. Thus we do not include an interaction term into our models as we see that it is not clarifying the effects much in SFA context. We emphasize that in this paper the

results of the SFA model are only reported in order to show that the misinterpretation of the coefficients $\boldsymbol{\theta}$ is indeed a relevant concern in the present context.

StoNED estimator

To avoid the problem noted in the previous section and to relax the restrictive functional form assumption of the production function, we also apply a more flexible semi-nonparametric specification. Recently, Johnson and Kuosmanen (2011) introduced a more flexible semi-nonparametric model:

$$\ln y_i = \ln \phi(\mathbf{x}_i) + \boldsymbol{\delta}'\mathbf{z}_i + \tilde{\varepsilon}_i, \tag{6}$$

where ϕ is an increasing and concave production function with an unspecified functional form. The composite disturbance term $\tilde{\varepsilon}_i$ contains the inefficiency and noise components, similar to the SFA model. To estimate Equation (6), Johnson and Kuosmanen (2011) develop a semi-nonparametric StoNED estimator where the effects of contextual variables are simultaneously estimated with the production function. The estimator can be seen as an extension of Kuosmanen and Johnson (2010) and Kuosmanen and Kortelainen (2012) to the case where contextual variables are taken explicitly into account. Kuosmanen and Kortelainen (2012) proposed the StoNED estimator as an amalgam of SFA and the nonparametric *Data Envelopment Analysis* (DEA).[6] Both SFA and DEA can be obtained as restricted special cases of the StoNED model. Note that Equation (6) reduces to (3) if the Cobb-Douglas functional form is imposed on ϕ.

The StoNED estimator is obtained as the optimal solution to the following least squares problem:

$$\min_{\alpha,\boldsymbol{\beta},\boldsymbol{\delta},\phi,e} \sum_{i=1}^{n} e_i^2 \tag{7}$$

subject to

$$\begin{aligned}
&\ln y_i = \ln \phi_i + \boldsymbol{\delta}'\mathbf{z}_i + e_i && \forall i = 1,\ldots,n \\
&\phi_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i && \forall i = 1,\ldots,n \\
&\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i && \forall h,i = 1,\ldots,n \\
&\boldsymbol{\beta}_i \ge \mathbf{0} && \forall i = 1,\ldots,n
\end{aligned}$$

In the problem (7), we are minimizing the sum of squared residuals e_i, which we will henceforth refer to as StoNED residuals. The first constraint is the empirical counterpart to Equation (6). The parameter ϕ_i is the estimated frontier output. The estimation of the nonparametric production function is based on the tangent hyperplanes defined by the second set of constraints. Concavity of the estimated frontier is ensured by the third set of constraints, which can be interpreted as the Afriat

inequalities (Afriat, 1967; 1972; see Kuosmanen, 2008, for a detailed discussion). Monotonicity of the production function follows from the non-negativity of $\boldsymbol{\beta}$ coefficients. The parameters that characterise the frontier ($\phi, \alpha, \boldsymbol{\beta}$) differ across countries, whereas the effects of contextual variables are common to all countries.

We examine the gamble hypothesis by empirically testing if contextual variables **z** can explain heteroskedasticity in the OLS and StoNED residuals. We apply the standard econometric approach to testing heteroskedasticity where the squared residuals are regressed on the variables associated with heteroskedasticity (White, 1980; Greene, 2003). Specifically, in our empirical application we regress the squared OLS and StoNED residuals on the contextual variables **z** according to the following equation

$$e_i^2 = \gamma_0 + \boldsymbol{\gamma}'\mathbf{z}_i + \tilde{v}_i, \tag{8}$$

where the random variable $\tilde{v}_i$ is the usual disturbance term of the linear regression model. Parameters $\boldsymbol{\gamma}$ represent the effects of variables **z** on heteroskedasticity. We can estimate Equation (8) by OLS and test the statistical significance of coefficients $\boldsymbol{\gamma}$ by using the conventional methods. In both OLS and StoNED approaches, the parameters $\boldsymbol{\delta}$ represent the effects of contextual variables on the expected level of productivity, whereas the parameters $\boldsymbol{\gamma}$ capture the variance effects.

IV. DATA AND VARIABLES

The empirical analysis uses two distinct cross sections from years 1990 and 2010. The first dataset from 1990 mimics the data used by Méon and Weill (2005). Using a comparable data, we examine the effects of corruption on the expected value and variance of TFP separately in order to test the grease, grit, and gamble hypotheses. In addition, to subject these hypotheses against the test of time, we use another cross section of data from year 2010, the latest year currently available to us. To fully ensure that the results between the two time periods are comparable, both of the datasets contain the same set of 80 countries. The variables in both datasets are briefly described next. The countries included in the study are listed in Appendix. The full data set and more detailed descriptions of the definitions and sources of data are provided by request from the authors.

Corruption indicators

In the quantitative corruption research at the macro level, the WGI and CPI indices are the two standard and most commonly used measures of corruption. In this study we consider the both indices as alternative measures of corruption. For the dataset of 1990, we only use the WGI index as the CPI index for the same number of countries

was not available. For the cross section of 2010 we use both WGI and CPI, which allows us to study the robustness of our results with respect to alternative corruption measures.[7]

For the sake of comparability, the original WGI and CPI indices have been rescaled to the interval [0,10] such that the sample minimum is 0 and the sample maximum is 10. For both indices, the value 10 refers to the lowest level of (perceived) corruption. In the 1990 data set we use the WGI index numbers from the year 1996 because this is the earliest year for which WGI indicator is available. However, the levels of corruption are generally considered to be stable over time (for example, McAdam & Rummel, 2004), and the corruption indices used in the present study support this view. Table 1 presents the summary statistics and correlation coefficients for the rescaled indices. All the indices are very highly positively correlated, as expected. We can see that the WGI indices from 1996 and 2010 have a very high positive correlation, which confirms its stability over time.

Table 1: Summary statistics and correlations of the corruption indices

	Mean	St. dev.	Median
WGI_1990	5.11	2.80	3.97
WGI_2010	4.06	3.04	2.74
CPI_2010	3.66	3.27	2.12
Correlation matrix	WGI_1990	WGI_2010	CPI_2010
WGI_1990	1		
WGI_2010	0.936	1	
CPI_2010	0.944	0.992	1

Macroeconomic data

Following Méon and Weill (2005), the macroeconomic data for the output, the capital stock, and the labor input in the cross section of 1990 are based on Easterly and Levine (2001). The data for these variables are obtained from World Bank's Global Development Network Growth database. The data for human capital is from Barro and Lee (2000), and it is downloaded from Center for International Development. Human capital is measured by the total years of schooling in the working age population (+15 years of age), calculated as average years of schooling times +15 aged population. The output and the capital stock are in Billions of U.S. dollars (at the prices of 1985; currencies converted using the PPP exchange rates), the human capital is in Millions of years, and the labor input is in Millions of workers.

For the cross section of 2010, data of the total population, output, number of workers and capital are obtained from the latest edition of Penn World tables (edition 7.1). The real output per worker is based on the PPP conversion rates and the prices of 2005. The capital stock in each year is calculated from the Penn data using the

perpetual inventory method (see Caselli, 2005, for details). The human capital measure is the same as in the 1990 cross section; the average years of schooling is obtained from the latest update of Barro and Lee (2010) dataset. To calculate the +15 aged population, the share of +15 aged population from total population is obtained from the World Bank.

The constant returns to scale (CRS) production function is generally preferred as a benchmark technology in cross-country productivity comparisons (Färe et al., 1994; Moroney & Lovell, 1997). To impose CRS in the Cobb-Douglas production function, the standard approach is to use scaled output and input variables obtained by dividing each variable with one of the inputs. In our empirical estimations we use the labor input as the scaling factor. The use of the scaled variables also imposes CRS in the StoNED method introduced in Section 2. Summary statistics of the scaled output and input variables used in the estimations are presented in Table 2.

Table 2: Summary statistics of the macroeconomic variables

	Mean	St. dev.	Min.	Max.
Output/Worker 1990	13310	10521	1107	36771
Capital/Worker 1990	29088	28746	288	103450
Hum.Cap./Worker 1990	10.37	5.01	0.80	21.47
Output/Worker 2010	31363	27643	606	101094
Capital/Worker 2010	90169	97848	375	305171
Hum.Cap./Worker 2010	13.81	4.95	2.07	23.35

Other contextual variables

In addition to the two corruption indices (WGI, CPI), we also control for three other contextual variables, following the specification by Méon and Weill (2005, 2010). The full vector of contextual variables is

$$\mathbf{z} = (Corruption, Openness, Ethnicity, Latitude) .$$

Corruption refers to the WGI and CPI indices. The other three contextual variables are the following.

Openness refers to openness to trade, measured as the total trade (imports plus exports) divided by the GDP. For the 1990 data, this variable is adopted from World Bank's Global Development Network Growth database. For the cross section of 2010, we use the figures from Penn Tables 7.1. We use the total trade to account for the possibility of export-led growth (Mookerjee, 2006; Wagner, 2007).[8] While evidence regarding the effects of trade openness on the economic performance is somewhat mixed (for example, Yanikkaya, 2003; Lee, Ricci & Rigobon, 2004), the effects seems to be positive in the presence of other favorable conditions such as good

institutions (Baldwin, 2003). Controlling openness is also justified on the grounds that it may influence the prevalence of corruption (Ades & Di Tella, 1999; Pellegrini & Gerlagh, 2004).

Ethnicity refers to the indicator of social, ethnic, and cultural fractionalization of the population. It is generally important to control for the ethnical fractionalization since fractionalised countries are often perceived to have decreased quality of government (La Porta et al., 1999). In the 1990 cross section, we use the *ELF* indicator by Roeder (2001). *ELF* is a proxy for the ethnolinguistic fractionalization, defined as the probability that two randomly drawn individuals in a country speak different native languages. The *ELF* indices are from the year 1985, as this is the latest year available in Roeder's dataset. Like corruption, ethnical fractionalization can be shown to be relatively stable over time. In the 2010 cross section, we use the *ETH* indicator from Alessina et al. (2003). While *ELF* index focuses on the linguistic aspect of fractionalization, the *ETH* indicator is more general in the sense that it also considers racial aspects.

Finally, *Latitude* is the absolute value of the average geographic latitude of the country, obtained from the OpenData webpage by Socrata Inc. The use of the absolute value implies that countries located North or South of the equator are treated symmetrically. The role of the latitude as a control variable is based on Sachs (2001), who found that tropical countries are generally on the lower level of development.

For the sake of comparability, our specification of contextual variables follows that of Méon and Weill (2005) as closely as possible. Of course, further contextual variables could be considered, but we prefer to leave this as a question for future research. For example, one could try to control for other aspects of governance, and also the interactions between governance and corruption (see, e.g., Méon and Weill, 2010). In this study, however, we focus on estimating the total effect of corruption, which includes the indirect effects of corruption that run through other governance aspects (see, e.g., Doucouliagos and Ulubaşoğlu, 2008, and Campos, Dimova, and Saleh, 2010, for further discussion on the direct and indirect effects). Thus, our estimates on the productivity effect can be interpreted to represent an upper bound of the direct effect as our estimates do not make distinction between the direct and indirect effects.

V. EMPIRICAL RESULTS

This section presents the results from empirical estimations. We first report the results briefly, following the same order as in Section 2. The baseline estimations with OLS are presented in Section (a), which are then followed by the SFA results in Section (b).

The corresponding StoNED results are presented in Section (c). In the Section (d) results are compared and discussed in more detail.

OLS results

The OLS estimates of the coefficients ($\boldsymbol{\delta}$,$\boldsymbol{\beta}$,α) of the baseline model (1) are reported in Table 3.[9] All three models yield good empirical fit, and the coefficients of the production factors are positive as expected. Our main interest is in the coefficients of the corruption indices. In all three models considered, find that the corruption indices have a positive and significant effect on the expected value of productivity shock. Recall that the high values of indices indicate low corruption, so the positive sign of the coefficient suggests that corruption has a negative effect on the expected productivity shock. In conclusion, the OLS estimates uniformly support the grit on the wheels hypothesis, and there is no support for the grease hypothesis.

Table 3: OLS estimates – the expected levels (δ)

DEP. VAR: ln(y/L)	WGI_1990	WGI_2010	CPI_2010
Corruption	0.047**	0.021*	0.023**
	(0.023)	(0.013)	(0.011)
Openness	-0.001	-0.00006	-0.00013
	(0.001)	(0.001)	(0.0005)
Ethnicity	-0.149	-0.093	-0.092
	(0.172)	(0.110)	(0.110)
Latitude	0.001	0.001	0.00024
	(0.004)	(0.002)	(0.002)
ln(K/L)	0.507***	0.740***	0.739***
	(0.060)	(0.041)	(0.038)
ln(H/L)	0.129	0.170*	0.172*
	(0.089)	(0.089)	(0.089)
Intercept	3.804***	1.397***	1.426***
	(0.480)	(0.305)	(0.295)
R^2	0.908	0.974	0.975

*** 1% significance, ** 5% significance, * 10% significance
White's heteroscedasticity robust standard errors in parentheses.

Consider the gamble hypothesis next. Table 4 reports the estimated coefficients ($\boldsymbol{\gamma}$) of Equation (8) where we regress the squared OLS residuals on the contextual variables $\mathbf{z}$. In all three models, the corruption index has a significant negative effect on heteroskedasticity. The degree of significance decreases as we move from left to right in the columns of Table 4: we discuss this point in Section (d) below. In conclusion, we find that corruption is associated with a higher variance of productivity as all the coefficients are consistently negative and statistically significant. The results of our heteroskedasticity tests support the gamble hypothesis.

Table 4: OLS estimates – the variance effects (γ)

DEP.VAR: e^2	WGI_1990	WGI_2010	CPI_2010
Corruption	-0.032***	-0.007**	-0.006*
	(0.011)	(0.004)	(0.003)
Openness	0.001	0.0002	0.00017
	(0.00044)	(0.00014)	(0.00015)
Ethnicity	-0.050	-0.011	-0.011
	(0.081)	(0.032)	(0.032)
Latitude	0.002	0.00018	0.00008
	(0.002)	(0.001)	(0.001)
Intercept	0.197***	0.051*	0.047
	(0.074)	(0.028)	(0.029)
R^2	0.149	0.092	0.073

*** 1% significance, ** 5% significance, * 10% significance
Standard errors in parentheses.

SFA results

It worth to emphasise again that the SFA model considered below is not suitable for testing the grease, grit, or gamble hypotheses as the effects of corruption on the expected value and variance of productivity are indistinguishable in this model. We report the empirical results of the SFA model for the sake of comparison, to illustrate that the misinterpretation of the SFA estimates is indeed a concern in the present context.

Table 5 presents the parameter estimates of the SFA model stated in Equation (4). The parameter estimates of the production factors are again positive; the only notable difference to the OLS coefficients is that the coefficient of the human capital becomes significant the WGI 1990 model. The stochastic frontier specification is supported by the LR test, which indicates that the parameters of the inefficiency model are jointly significant.

In all three model specifications considered, the signs of the corruption index are systematically negative. However, the coefficient of the corruption index is statistically significant only in the WGI 1990 model. Note that in the SFA specification the corruption indices influence inefficiency. Therefore the negative sign of the corruption coefficient in SFA is in line with the positive sign of the OLS coefficient that represents the direct marginal effect of corruption on output. The negative sign of the corruption coefficient in SFA also conforms to the results reported in earlier SFA studies (Méon & Weill, 2005; 2010)

As discussed in Section 3, it might be tempting to interpret the negative coefficients of corruption in the SFA model as support for the grit hypothesis. However, it is equally possible that the negative coefficients of corruption are driven

by heteroskedasticity. Comparison with the OLS results obtained using the same Cobb-Douglas production function specification suggests that corruption influences both the expected level and the variance of productivity. Since in SFA both effects are captured by the same coefficient, there is a major risk of confusion.

Table 5: SFA estimates

DEP. VAR. ln(y/L)	WGI_1990	WGI_2010	CPI_2010
Corruption	-0.156***	-0.086	-0.098
	(0.058)	(0.072)	(0.079)
Openness	0.001	0.001	0.002
	(0.002)	(0.002)	(0.002)
Ethnicity	0.211	0.211	0.235
	(0.337)	(0.321)	(0.318)
Latitude	-0.008	-0.001	2.6e-4
	(0.008)	(0.006)	(0.006)
Intercept	0.960***	0.152	0.064
	(0.327)	(0.404)	(0.437)
ln(K/L)	0.440***	0.733***	0.732***
	(0.026)	(0.027)	(0.027)
ln(H/L)	0.155**	0.183**	0.183**
	(0.076)	(0.076)	(0.077)
Intercept	5.024***	1.720***	1.724
	(0.137)	(0.283)	(0.280)
Sigma-squared	0.182	0.071	0.069
Log-likelihood	-3.165	21.217	21.914
LR test	40.534***	14.090**	15.484**

*** 1% significance, ** 5% significance, * 10% significance
Standard errors in parentheses.

StoNED results

To assess the robustness of the previous results to the potentially restrictive parametric specification of the Cobb-Douglas production function, we consider a more flexible semi-nonparametric estimator that allows the output elasticities of production factors differ across countries. The StoNED estimator is obtained by solving the least squares problem presented in Equation (7). Recall that the coefficients of the nonparametric production function (α_i,$\boldsymbol{\beta}_i$) refer to the tangent hyperplanes of an unspecified functional form, and as such are not comparable with the coefficients of the Cobb-Douglas production function. Moreover, since the coefficients (α_i,$\boldsymbol{\beta}_i$) are country-specific, and not necessarily unique, we do not report them in this paper. Instead, we resort to a graphical illustration of the estimated production function in Figure 2 below. But first, let us examine the effects of the contextual variables (coefficients $\boldsymbol{\delta}$), which are presented in Table 6.

The corruption indices have systematically positive signs in all three samples, analogous to the OLS estimates. Interestingly, the coefficient of corruption is

not significant in the WGI 1990 sample, but it is significant in both the WGI and CPI samples for the year 2010. The results of the StoNED estimator support the view from the OLS regression: there is evidence in favor of the grit on the wheels hypothesis, but no indication whatsoever in favor of the grease hypothesis.

The coefficients of determination (R^2) reported on the bottom row of Table 6 refer to the StoNED model as a whole, including the input variables (not reported in the table). These coefficients are higher than the corresponding OLS statistics, which indicates that the StoNED estimator has a slightly better empirical fit. The skewness statistics refer to the skewness of the StoNED residuals. This statistic is expected to be negative in the case of a production function when there is asymmetric inefficiency in the disturbance term. The negative skewness statistics suggest that there is inefficiency present in all three samples. However, we do not report or discuss the country-specific inefficiency estimates, as this falls beyond the scope of the present study.

Table 6: StoNED estimates – the expected levels (δ)

DEP. VAR. $\ln(y/\phi)$	WGI_1990	WGI_2010	CPI_2010
Corruption	0.012	0.024**	0.024***
	(0.018)	(0.009)	(0.009)
Openness	-0.00033	0.00002	-0.000011
	(0.001)	(0.00047)	(0.00046)
ELF	-0.166	-0.027	-0.028
	(0.145)	(0.099)	(0.099)
Latitude	-0.00041	0.001	0.001
	(0.003)	(0.002)	(0.002)
Intercept	0.000	0.000	0.000
	(0.137)	(0.081)	(0.080)
R^2	0.931	0.978	0.978
Skewness	-0.590	-0.328	-0.352

*** 1% significance, ** 5% significance, * 10% significance
White's heteroscedasticity robust standard errors in parentheses.

Consider next the gamble hypothesis. We estimate Equation (7) by OLS, regressing the squared StoNED residuals on the contextual variables. The estimated coefficients and their standard errors are reported in Table 7. The coefficients of corruption are systematically negative throughout all three models considered. The negative coefficient for corruption implies that the variance of StoNED residuals decreases as the corruption indices increase (that is, corruption decreases). In other words, highly corrupted countries have a larger variation in their productivity levels than less corrupted countries. In contrast to the effects on the expected level, the coefficient in WGI 1990 sample is statistically significant at 5% significance level, whereas in the 2010 samples the coefficients are closer to zero and significant only at

10% significance level. In our interpretation, corruption is clearly a risk factor in the 1990 sample, with little effect on the level of productivity, but in the data from 2010, the variance effect has notably decreased and the effect on the expected level is more dominant.

Table 7: StoNED estimates – the variance effects (γ)

DEP.VAR: e^2	WGI_1990	WGI_2010	CPI_2010
Corruption	-0.022**	-0.006*	-0.005*
	(0.008)	(0.003)	(0.003)
Openness	0.00054	0.0002	0.00019
	(0.00035)	(0.00013)	(0.00013)
ELF	-0.103	-0.001	-0.00012
	(0.064)	(0.029)	(0.029)
Latitude	0.001	0.00014	0.00014
	(0.001)	(0.001)	(0.001)
Intercept	0.174***	0.038**	0.033**
	(0.058)	(0.025)	(0.026)
R^2	0.139	0.090	0.083

*** 1% significance, ** 5% significance, * 10% significance
Standard errors in parentheses.

To get a visual impression about heteroskedasticy, in Figure 1 we have plotted the StoNED residuals against the WGI corruption index. The upper panel in the figure is the model with 1990 data, whereas in the left panel we have the obtained residual using the 2010 data. This figure illustrates that the variation in residuals is notably higher in the lower end of the corruption index, where highly corrupted countries are.

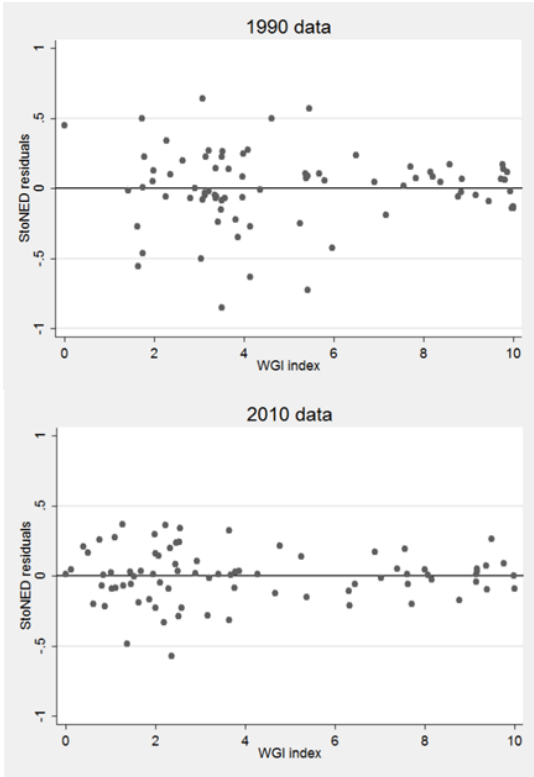

Figure 1: StoNED residuals vs. WGI index

Finally, it is worth to examine the shape of the estimated StoNED production function graphically. The upper panel of Figure 2 illustrates the production function for year 1990 and the lower panel for year 2010. Whereas in OLS and SFA cases (see the online appendix for graphical illustrations) the Cobb-Douglas production function imposes the substitution elasticity between inputs to be equal to one by construction, in the case of the StoNED frontier, the range of substitution elasticity is from zero to infinity. According to Figure 2, when the capital intensity is low in terms of both human and physical capital, the substitution possibilities seem to be rather limited. In our interpretation, this reflects the *technology-skill mismatch* in low development countries. Acemoglu and Zilibotti (2001) emphasise this mismatch as a source of the productivity differences at different levels of development.[10] The technology-skill mismatch occurs because the physical capital itself does not increase production if there are no skills to utilise it. Conversely, the human capital alone does not increase the output without a sufficient level of physical capital. This implies that the production function must exhibit the Leontief type fixed proportions structure at the low level of development. The substitution elasticity of human and physical capital increases at the higher levels of capital intensity. This suggests that it is easier to match the technology with the necessary skills or vice versa in more developed

countries. As the capital intensity continues to increase, the slope of the production function becomes flatter due to the diminishing returns. In our view, the shape of the estimated StoNED production function thus has a meaningful economic interpretation.

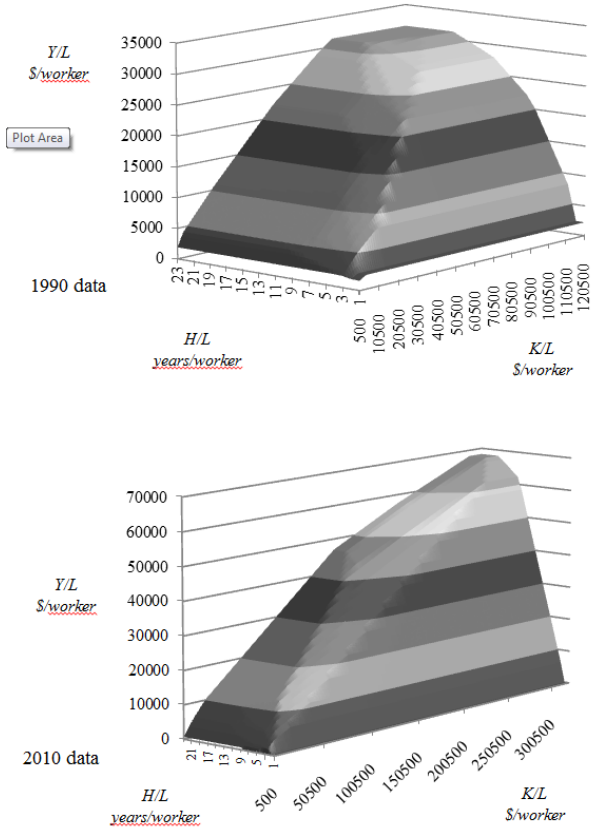

Figure 2: Estimated StoNED production function (WGI)

Comparison of the production functions of years 1990 and 2010 reveals clear signs of economic growth. The output levels have increased considerably during the twenty year period, but also the amount of physical capital has increase as the latter picture extends further in *K/L* axis. This also explains why the 2010 picture looks somewhat stretched. Indeed, if we restrict the capital per worker in 2010 to the observed range of *K/L* in year 1990, the two graphs would appear almost identical. Thus the overall shape of the production function has not changed, but rather we see capital deepening in our results.

Discussion of the results

By drawing a clear distinction between the effects to the expected level and the variance, our results offer new insights to the adverse effects of corruption. We find that besides the effect of corruption on the expected level of productivity, there is the

added effect of increase in the risk. Previous studies that focus only on the expected level ignore the corresponding increase in the risk, which may be even more harmful consequence of corruption.

Both OLS and StoNED estimates indicate that corruption has a significant positive effect on heteroskedasticity: corruption increases the variance of productivity. Thus, we find significant support for the gamble hypothesis. The result is robust to the functional form assumption (Cobb-Douglas used in OLS versus non-parametric production function of StoNED), the time period (1990 sample versus 2010 sample), and the choice of the corruption index (WGI versus CPI in 2010). Note that functional form however has notable effect for the level effects of corruption when we consider the 1990 cross section. The level effect is highly significant with OLS, but turns insignificant when using StoNED. We do not see equally notable changes with the heteroskedasticity effects. Thus the risk effect seems somewhat more robust.

When comparing the relative importance of level and variance effects of corruption, Both OLS and StoNED estimates show that the heteroskedasticity effect dominates in the 1990 cross section. In the 2010 cross section the effect on the expected level becomes more pronounced. In our interpretation, this is a natural consequence of economic growth and institutional convergence across countries. As countries improve their institutions over time, countries converge towards a more uniform growth path. Hence the heteroskedasticity that prevails at the lower levels of development will gradually vanish, and the effect on the expected level becomes more visible. In other words, the institutions become increasingly important at higher levels of economic development (see Glaeser et al., 2004; Aidt et al., 2008, for further discussion).

Let us consider next the above result in the light of the SFA results. A major shortcoming of the SFA model considered above is that it cannot distinguish the effects on the expected level and the variance. In the SFA results, the negative coefficient of corruption implies that corruption increases the variance of the inefficiency term (see Equation (5)), conforming to the OLS and StoNED results. By comparison with the OLS and StoNED results, it seems that the SFA coefficient of corruption is very sensitive to the heteroskedasticity effect. Indeed, the SFA coefficient of corruption is highly significant in the 1990 cross section where the variance effect is dominant in OLS and StoNED. In contrast, in the 2010 cross section, where the level effects are more dominant, the SFA coefficient of corruption is insignificant. These observations suggest that partial support for the grease hypotheses reported in the literature, obtained with similar SFA models as the one considered in this paper, may in fact result from a misinterpretation of the SFA coefficient that captures the heteroskedasticity effect.

The empirical support on the gamble hypothesis of corruption suggests a new perspective to corruption as a source of macro risk in the fragile stages of development. Countries at their early stage of economic development seem to differ in their equilibrium levels of corruption, and consequently have different growth paths. The probabilistic interpretation of corruption as a risk factor can help to understand why some highly corrupted countries have managed to achieve relatively high level of productivity. A high variance means that large deviations to both positive and negative direction from the mean are likely. The gamble hypothesis allows for the possibility that, with some good fortune, it is possible that a country achieves a relatively high level of productivity despite a high prevalence of corruption. In our interpretation, the mixed empirical evidence supporting the grease or the grit hypotheses can at least partly be due to the variance effect of corruption.

VI. CONCLUSIONS

Previous empirical studies (Méon and Sekkat, 2005; Méndez & Sepúlveda, 2006; Aidt, Dutta & Sena, 2008; Méon & Weill, 2010) rather unanimously show that the effect of corruption is not uniform across countries and over time, but the effect depends on other factors such as institutions and economic freedom. These studies have however focused predominantly on estimating the effect of corruption on the expected level of productivity or its growth rate. In this study we argue that corruption also affects the variance of productivity, which leads us to propose a new gamble hypothesis of corruption. Gamble hypothesis contributes to the literature by introducing a new perspective: the risk. According to this probabilistic hypothesis, corruption increases the variance of productivity, and can hence be seen as a source of macro risk. The gamble hypothesis predicts a large dispersion of productivity levels at high levels of corruption and a convergence to a more uniform productivity levels at low levels of corruption. Note that high variance involves both positive and negative risk. It is possible that a highly corrupted country achieves a high level of productivity, but the probability of low productivity is also high.

Our empirical results indicate that corruption has a significant positive effect on the variance of productivity, supporting the gamble hypothesis. Especially at the higher corruption levels, we see large variations on how corruption affects to the productivity of countries. This result is robust to the choice of the corruption index, the estimation method, and the time period. We also find some support for the grit in the wheels hypothesis, which suggests that corruption has a negative effect on the expected level of productivity, particularly in the latest cross section from the year 2010. We do not find evidence to support the idea that corruption is directly beneficial for economic performance, implying that grease hypothesis can be refuted. At most,

the weak form of the grease hypothesis, as suggested by Méon and Weill (2010), can be seen somewhat compatible with our results. Thus besides the expected effect of better institutions positively contributing to the level of productivity, there is the added benefit through the lower variance of productivity, that is, a lower macro risk. Decreasing risk can be an important motivation to fight corruption, particularly for risk averse governments.

Since our results suggest that corruption increases the macro risk, the policy against corruption must similarly involve uncertainty. Therefore, it is hard to identify and quantify the exact effects of anti-corruption policies as these effects vary depending on the external conditions. We see this as an important lesson for the policy makers. Consequently, the cost-benefit analyses of anti-corruption policy actions should carefully take into account the institutional setting of a country. The optimal policy measures to battle corruption are likely to vary across countries. Recall the points by Shleifer and Vishny (1993) that were briefly discussed earlier: regardless of whether we observe corruption in decentralized or centralized form, the fixes are unlikely to be the same. In the latter case, the rules of corruption are likely to be rather clear, and hence changing the rules can be an effective way to curb corruption. In the former case, however, to mitigate the harmful effects of corruption, some rules need to be established in the first place.

Of course, as always, some assumption and limitations of our study should be kept in mind when interpreting our results. These limitations also offer new avenues to extend our analysis further. Firstly, we could extend the set of control variables to include indicators of other aspects of governance or interactions between the variables, as already noted. The parametrization assumed in our study assumes that corruption has a linear effect on the expected value and variance of the productivity shock. While this parametrization allows us easily to distinguish the heteroskedasticity effects from the level effects, it may ignore potential further nonlinear effects of corruption on the level of economic output, which are likely to be characterized by the interactions between the variables (see for example, Tan, 2010). Thus it is possible that violations of the linearity assumption appear as heteroskedasticity, and conversely, heteroskedasticity may appear empirically as nonlinearities. In future, we could for example study how the degree of centralization of institutions is related to the effect of corruption as different levels of centralization have been argued to more prone to corruption than others.[11] If different levels of centralization imply different levels and types of corruption, it is likely that the effect of corruption varies accordingly.

Secondly, we have only examined the effect of corruption to the overall error, which consist both noise and inefficiency. It would be interesting to try to

attribute heteroskedasticity to these subcomponents of the overall error separately. Lastly we acknowledge that reverse causality may have some bearing on the results considering the level effects. That is, we have assumed that causality runs only from corruption to productivity, not the other way around. However, we argue that reverse causality cannot explain the heteroskedasticity effects, which we consider as the main finding of this study. The variance of productivity depends on the level of corruption irrespective of whether corruption drives productivity, or vice versa.

ENDNOTES

1. The terms environmental variables and *z*-variables are also commonly used in the literature.
2. The distinction between *productivity* and *efficiency* is worth clarifying. Productivity is defined as the ratio of output to input (or the index of inputs). Output efficiency is defined as the ratio of the observed output and the maximum output. Productivity change can be decomposed into the components of efficiency change and technical progress, and possibly some other components (see for example, Färe et al., 1994; Kuosmanen & Sipiläinen, 2009).
3. For a more extensive treatment of SFA, we refer to the book by Kumbhakar and Lovell (2000).
4. StoNED is an abbreviation of *Stochastic semi-Nonparametric Envelopment of Data* (Johnson & Kuosmanen, 2011, footnote 3).
5. This issue is related to the notion of *scaling property* (Wang & Schmidt, 2002; Alvarez et al., 2006) in SFA models. The KGM model does not satisfy the scaling property, as Wang and Schmidt (2002) point out.
6. Nonparametric DEA applies a similar axiomatic nonparametric frontier, but in contrast to StoNED, DEA neglects the noise term altogether. We consider this a major limitation in the present context. Further, regressing DEA efficiency estimates on contextual variables is known to be problematic even if one assumes away noise (see Simar and Wilson, 2007, for a sharp critique of two-stage DEA). The problems concern particularly the statistical inferences. For these reasons, we do not consider DEA appropriate for estimating the effect of corruption on the level of productivity, let alone its variance.
7. In the 1990 data, the CPI index is available for a subset of 48 countries. Using this subset of 48 countries, the results are rather similar than with the larger sample. These results with the sample of 48 countries can be found from the online appendix.

8. The causality from exports to productivity is subject to debate. For example, Yamada (1998) does not find strong evidence for the causality, but mentions that the degree of causality may differ across countries (see also Konya, 2004).
9. Empirical analysis was conducted using the following software: Stata 11 (OLS and SFA) see and GAMS using MINOS solver (StoNED).
10. Murphy, Shleifer, and Vishny (1991, 1993) suggest rent-seeking and corruption as possible explanations for the misallocation of talent.
11. For example Gerring and Thacker (2008) argue that centralized political systems, reduce the prevalence of corruption largely because fragmented systems create more opportunities for corruption and the path of accountability is often more clearer in a centralized system (see also Tavis, 2007). In contrast, centralized systems have been criticized on the basis of reducing the amount of checks and balances, and competition among bureaucrats, thus inducing more chances on corruption. Thus any micro-level actions against corruption are ineffective in the decentralized case and the system should be centrally restructured. In contrast, in the centralized case the benefits of centralization may be questioned. Note that by centralized system Gerring and Thacker (2008) refer to a system that has a strong national government and which follows parliamentarism. In such systems the electoral threat should impose some checks on political corruption. On the opposite, one can expect that dictatorial regimes are more likely to exhibit risk seeking corrupt behavior than democratically elected governments as dictatorial regimes are not subject to electoral threat (Quinn & Woolley, 2001).

REFERENCES

Acemoglu D., & Zilibotti, F. (2001). Productivity differences. *Quarterly Journal of Economics* 116(2), 563-606.

Ades A., & Di Tella, R. (1999). Rents, competition, and corruption. *American Economic Review*, 89(4), 982-993.

Aidt, T., Dutta, J., & Sena, V. (2008). Governance regimes, corruption and growth: Theory and evidence. *Journal of Comparative Economics*, 36, 195-220.

Adkins, L.C., Moomaw, R.L., & Savvides, A. (2002). Institutions, freedom, and technical efficiency. *Southern Economic Journal*, 69(1), 92-108.

Afriat, S.N. (1967). The construction of a utility function from expenditure data. *International Economic Review*, 8, 66-77.

Afriat, S.N. (1972). Efficiency estimation of production functions. *International Economic Review*, 13(3), 568-598.

Alessina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., & Wacziarg, R. (2003). Fractionalization. *Journal of Economic Growth*, 8, 155-194. (The data is downloadable from: http://www.nsd.uib.no/macrodataguide/set.html?id=16andsub=1 , Last accessed 20.12.2012)

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

Assane, D., & Grammy, A. (2003). Institutional framework and economic development: international evidence. *Applied Economics*, 35, 1811-1817.

Bardhan, P. (1997). Corruption and development: A review of issues. *Journal of Economic Literature*, 35(5), 1320-1346.

Bardhan, P. (2006). The economist's approach to the problem of corruption. *World Development*, 34(2), 3342-348.

Barro, R.J., & Lee, J.-W. (2000). International data on educational attainment: Updates and implications. Harvard University, CID Working Paper No. 42, April 2000. Downloadable at http://www.cid.harvard.edu/ciddata/ciddata.html (Last accessed 2.8.2012).

Barro, R.J., & Lee, J.-W. (2010). A New Data Set of Educational Attainment in the World. Downloadable at http://www.barrolee.com/ (Last accessed 20.12.2012).

Baldwin, R.E. (2003). Openness and growth: What's the empirical relationship? NBER Working Paper No. 9578.

Battese, G.E., & Coelli. T.J. (1995). A model for technical inefficiency effects in a stochastic production frontier for panel data. *Empirical Economics*, 20(2), 325-332.

Beenstock, M. (1979). Corruption and development. *World Development*, 7, 15-24.

Blackburn, K., & Forgues-Puccio, G.F. (2009). Why is corruption less harmful in some countries than in others. *Journal of Economic Behavior and Organization*, 72, 797-810.

Campos, N. F., Dimova, R., and Saleh, A. (2010). Whither corruption? A quantitative survey of the
literature on corruption and growth. *IZA Discussion paper series*, 5334. Downloadable at: http://ftp.iza.org/dp5334.pdf (Last accessed 12.2.2014).

Cavalcanti, T.V.V., & Álvaro, A.N. (2005). Institutions and economic development: How strong is the relation? *Empirical Economics*, 30, 263-276.

Caselli, F. (2005). Accounting for cross-country income differences. In P. Aghion, and S.F. Durlauf (Eds), *Handbook of Economic Growth*, Volume 1.A (pp. 679-741). Elsevier.

Cherchye, L., & Moesen, W. (2003). Institutional infrastructure and economic performance: Levels
versus catching up and frontier shifts. Public Economics Working Paper Series 03.14. K.U. Leuven.

Doucouliagos, H. & Ulubaşoğlu, M. A. (2008). Democracy and Economic Growth: A Meta-Analysis. *American Journal of Political Science*, 52 (1), 61-83.

Durlauf, S., & Johnson, P.A. (1995). Multiple regimes and cross-country growth behaviour. *Journal of Applied Econometrics*, 10, 365-384.

Easterly, W., & Levine, R. (2001). It's not factor accumulation: Stylized facts and growth models. *World Bank Economic Review*, 15(2), 177-219.

Ehrlich. I., & Lui, F.T. (1999). Bureaucratic corruption and endogenous economic growth. *The Journal of Political Economy*, 107(6), S270-S293.

Everhart, S., Martinez-Vasquez, J., & McNab, R.M. (2009). Corruption, governance, investment and growth in emerging markets. *Applied Economics*, 41, 1579-1594.

Färe, R., Grosskopf, S., Norris, M., & Zhang, Z. (1994). Productivity growth, technical progress and efficiency change in industrialized countries. *American Economic Review*, 84(1), 66-83.

Gerring, J. and Thacker, S. C. (2004). Political institutions and corruption: The role of unitarism and parliamentarism. *British Journal of Political Science*, 34, 295-330.

Glaeser, E.L., La Porta, R., Lopez-deSilanes, F., & Shleifer, A. (2004). Do institutions cause growth? *Journal of Economic Growth*, 9 (3), 271-303.

Greene, W.H. (2003). *Econometric Analysis*, 5th edition, Prentice Hall.

Haggard, S., & Tiede, L. (2011). The rule of law and economic growth: Where are we? *World Development*, 39(5), 673-685.

Hall, R.E., & Jones, C.I. (1999). Why do some countries produce so much more output per worker than others? *The Quarterly Journal of Economics*, 114(1), 83-116.

Hauner, D., & Kyobe, A. (2010). Determinants of government efficiency. *World Development*, 38(11), 1527-1542.

Herzfeld, T. & Weiss, C. (2003). Corruption and legal (in)effectiveness: an empirical investigation. *European Journal of Political Economy*, 19, 621-632.

Herzfeld, T. & Weiss, C. (2007). Corruption clubs: empirical evidence from kernel density estimates. *Applied Economics*, 39, 1565-1572.

Heston, A., & Kumar, V. (2008). Institutional flaws and corruption incentives in India. *Journal of Development Studies*, 44(9), 1243-1261.

Huntington, S.P. (1968). *Political order in changing societies*. Yale University Press, New Haven.

Jayasuriya, R., & Wodon, Q. (2005). Measuring and explaining the impact of productive efficiency on economic development. *The World Bank Economic Review*, 19(1), 121-140.

Johnson, A.L., & Kuosmanen, T. (2011). One-stage estimation of the effects of operational conditions and practices on productive performance: Asymptotically normal and efficient, root-n consistent StoNED method. *Journal of Productivity Analysis*, 36(2), 219-230.

Johnson, A.L., & Kuosmanen, T. (2012). One-stage and two-stage DEA estimation of the effects of contextual variables. *European Journal of Operational Research*, 220(2), 559–570.

Khan, M.H. (1996). The efficiency implications of corruption. *Journal of International Development*, 8(5), 683-696.

Konya, L. (2004). Export-led growth, growth-driven export, both or none? Granger causality on OECD countries. *Applied Econometrics and International Development*, 4(1), 73-94.

Kumar, S., & Russell, R. R. (2002). Technological change, technological catch-up, and capital deepening: contribution to growth and convergence. *American Economic Review*, 92 (3), 527-548.

Kumbhakar, S.C., Ghosh, S., & McGuckin, J.T. (1991). A generalized production frontier approach for estimating determinants of inefficiency in U.S. dairy farms. *Journal of Business and Economic Statistics*, 9, 279-286.

Kumbhakar, S.C., & Lovell, C.A.K. (2000). *Stochastic Frontier Analysis*. Cambridge University Press, New York.

Kuosmanen, T. (2008). Representation Theorem for Convex Nonparametric Least Squares. *Econometrics Journal*, 11, 308-325.

Kuosmanen, T., & Johnson, A.L. (2010). Data envelopment analysis as nonparametric least-squares regression. *Operations Research*, 58, 149-160.

Kuosmanen, T., Johnson, A. L., & Saastamoinen, A. (2014). Stochastic nonparametric approach to efficiency analysis: A unified framework. Forthcoming in J. Zhu (Ed.) Handbook on Data Envelopment Analysis Vol II, Springer.

Kuosmanen, T., & Kortelainen, M. (2012). Stochastic non-smooth envelopment of data: Semi-parametric frontier estimation subject to shape constraints. *Journal of Productivity Analysis* 38(1), 11-28.

Kuosmanen, T., & Sipiläinen, T. (2009). Exact decomposition of the Fisher ideal total factor productivity index. *Journal of Productivity Analysis* 31(3), 137-150.

Lambsdorff, J.G. (2003). How corruption affects productivity. *Kyklos* 56(4), 457-474.

Lambsdorff, J.G. (2005). Consequences and causes of corruption: What do we know from a cross-section of countries? Discussion paper, University of Passau, No. V-34-05, May 2005. Downloadable at http://www.icgg.org/corruption.research.html (Last accessed 28.10.2010).

La Porta, R., Lopez-deSilanes, F., Shleifer, A., & Vishny, R. (1999.) The quality of government. *Journal of Law, Economics and Organization*, 15(1), 222-279.

Lee, H.Y., Ricci, L.A., & Rigobon, R. (2004). Once again, is openness good for growth? *Journal of Development Economics*, 75(2), 451-472.

Leff, N.H. (1964). Economic development through bureaucratic corruption. *American Behavioral Scientist*, 8(3), 8-14.

Leys, C. (1965). What is the problem of corruption. *Journal of Modern African Studies*, 3(2), 215-230.

Mauro, P. (1995). Corruption and Growth. *The Quarterly Journal of Economics*, 101(3), 681-712.

Mauro, P. (1998). Corruption and the composition of government expenditure. *Journal of Public Economics*, 69(2), 263-279.

McAdam, P., & Rummel, O. (2004). Corruption: A Non-Parametric Analysis. *Journal of Economic Studies*, 31(6), 509-523.

Méndez, F., & Sepúlveda, F. (2006). Corruption, growth and political regimes: Cross country evidence. *European Journal of Political Economy*, 22, 82-98.

Méon, P-G., & Sekkat, K. (2005). Does corruption grease or sand the wheels of growth. *Public Choice*, 122, 69-97.

Méon, P-G., & Weill, L. (2005). Does better governance foster efficiency? An aggregate frontier analysis. *Economics of Governance*, 6, 75-90.

Méon, P.-G., & Weill, L. (2010). Is Corruption an Efficient Grease? *World Development*, 38(3), 244-259.

Mo, P-K. (2001). Corruption and growth. *Journal of Comparative Economics*, 29, 66-79.

Mookerjee, R. (2006). A meta-analysis of the export growth hypothesis. *Economics Letters*, 91(3), 395-401.

Moroney, J., & Lovell, C.A.K. (1997). The relative efficiencies of market and planned economies. *Southern Economic Journal*, 633(4), 1084-1093.

Murphy, K.M., Shleifer, A., & Vishny R.W. (1991). The allocation of talent: Implications for growth. *Quarterly Journal of Economics*, 106(2), 503-530.

Murphy, K.M., Shleifer, A., & Vishny, R.W. (1993). Why is rent-seeking so costly to growth? *American Economic Review*, 83(2) (Papers and Proceedings), 409-414.

Nye, J.S. (1967). Corruption and development: a cost-benefit analysis. *The American Political Science Review*. 61(2), 417-427.

OpenData, Average latitude and longitude of countries. Downloadable at: https://opendata.socrata.com/dataset/Country-List-ISO-3166-Codes-Latitude-Longitude/mnkm-8ram (Last accessed 29.1.2013)

Pellegrini, L., & Gerlagh, R. (2004). Corruption's effect on growth and its transmission channels. *Kyklos*, 57(3), 429-459.

Quinn, D.P., & Woolley, J.T. (2001). Democracy and national economic performance: the preference for stability. *American Journal of Political Science*, 45(3), 634-657.

Rock, M.T., & Bonnett, H. (2004). The comparative politics of corruption: accounting for the East Asian paradox in empirical studies of corruption, growth, and investment. *World Development*, 32(6), 999-1017.

Roeder, P.G. (2001). Ethnolinguistic fractionalization (ELF) Indices, 1961 and 1985. Downloadable at http://weber.ucsd.edu/~proeder/elf.htm (Last accessed 2.8.2012).

Sachs, J.D. (2001). Tropical underdevelopment. *NBER Working Paper Series*, Working Paper 8119, February 2001.

Salinas-Jiménez, M., & Salinas-Jiménez, J. (2007). Corruption, efficiency and productivity in OECD countries. *Journal of Policy Modeling*, 29, 903-915.

Salinas-Jiménez, M., & Salinas-Jiménez, J. (2010). Corruption and total factor productivity: level or growth effects? *Portuguese Economic Journal*, 10(2), 109-128.

Shleifer, A., & Vishny, R. W. (1993). Corruption. *The Quarterly Journal of Economics*, 108(3), 599-617.

Simar, L., & Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics*, 136, 31-64.

Tan, C.M. (2010). No one true path: uncovering the interplay between geography, institutions, and fractionalization in economic development. *Journal of Applied Econometrics*, 25, 1100-1127.

Tavis, M. (2007). Clarity of responsibility and corruption. *American Journal of Political Science*, 51(1), 218-229.

Transparency International, Corruption Perception Index, Downloadable at http://www.transparency.org/research/cpi/overview (Last accessed 29.1.2013).

Vial, V., & Hanoteu, J. (2010). Corruption, Manufacturing Plant Growth, and the Asian Paradox: Indonesian Evidence. *World Development*, 38(5), 693-705.

Wagner, J. (2007). Exports and productivity – comparable evidence for 14 countries. *LICOS Discussion Paper Series 192/2007,* Downloadable at: http://www.econ.kuleuven.be/licos/publications/dp/dp192.pdf (Last accessed 2.8.2012).

Wang, H., & Schmidt, P. (2002). One Step and Two Step Estimation of the Effects of Exogenous Variables on Technical Efficiency Levels. *Journal of Productivity Analysis*, 18, 129-144.

Weill, L. (2006). On the consistency of aggregate production frontiers. *European Journal of Operational Research*, 172, 326-333.

Weill, L. (2008). On the Inefficiency of European Socialist Economies: Relative to Developed and Developing Economies. *Journal of Productivity Analysis* 29, 79-89.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48(4), 817-838.

World Bank, Global Development Network Growth Database, Downloadable at http://econ.worldbank.org/ (Last accessed 2.8.2012).

World Bank, Worldwide Governance Indicators, Downloadable at http://info.worldbank.org/governance/wgi/index.asp (Last accessed, 7.1.2013).

Wyatt, G. (2003). Corruption, productivity, and socialism. *Kyklos*, 56, 223-244.

Yamada, H. (1998). A note on the causality between export and productivity: An empirical re-examination. *Economics Letters*, 61(1), 111-114.

Yanikkaya, H. (2003). Trade openness and economic growth: A cross-country empirical investigation. *Journal of Development Economics*, 72(1), 57-89.

APPENDIX

Countries included in the empirical estimations:

Algeria	France	Korea, Rep.	Sierra Leone
Argentina	Gambia	Malawi	Singapore
Australia	Germany	Malaysia	South Africa
Austria	Ghana	Mali	Spain
Bangladesh	Greece	Mauritius	Sri Lanka
Belgium	Guatemala	Mexico	Sweden
Bolivia	Guinea-Bissau	Mozambique	Switzerland
Brazil	Guyana	Netherlands	Syrian
Cameroon	Hungary	New Zealand	Thailand
Canada	Iceland	Nicaragua	Togo
Chile	India	Norway	Trinidad and Tobago
China	Indonesia	Pakistan	Tunisia
Colombia	Iran	Panama	Turkey
Cyprus	Ireland	Papua New Guinea	Uganda
Denmark	Israel	Paraguay	United Kingdom
Dominican Republic	Italy	Peru	United States
Ecuador	Jamaica	Philippines	Uruguay
Egypt	Japan	Poland	Venezuela
El Salvador	Jordan	Portugal	Zambia
Finland	Kenya	Senegal	Zimbabwe

Article 5

Antti Saastamoinen; Timo Kuosmanen. Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland. Forthcoming in Energy Economics, published online.
DOI: 10.1016/j.eneco.2014.04.016

ARTICLE IN PRESS
ENEECO-02775; No of Pages 12

Energy Economics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Energy Economics

journal homepage: www.elsevier.com/locate/eneco

Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland

Antti Saastamoinen *, Timo Kuosmanen

Aalto University School of Business, Runebergingkatu 22-24, 00100 Helsinki, Finland

ARTICLE INFO

Article history:
Received 26 September 2013
Received in revised form 21 March 2014
Accepted 23 April 2014
Available online xxxx

JEL classification:
L51
Q48
D24

Keywords:
Electricity distribution
Productivity and efficiency analysis
Regulation
Service quality

ABSTRACT

Electricity distribution is a prime example of local monopoly. In most countries, the costs of electricity distribution operators are regulated by the government. However, the cost regulation may create adverse incentives to compromise the quality of service. To avoid this, cost regulation is often amended with quality incentives. This study applies theory and methods of productivity analysis to model the frontier of service quality. A semi-nonparametric estimation method is developed, which does not assume any particular functional form for the quality frontier, but can accommodate stochastic noise and heteroscedasticity. The empirical part of our paper examines how underground cabling and location affect the interruption costs. As expected, higher proportion of underground cabling decreases the level of interruption costs. The effects of cabling and location on the variance of performance are also considered. Especially the location is found to be a significant source of heteroscedasticity in the interruption costs. Finally, the proposed quality frontier benchmark is compared to the current practice of Finnish regulation system. The proposed quality frontier is found to provide more meaningful and stable basis for setting quality targets than the average practice benchmarks currently in use.

1. Introduction

The last two decades have witnessed a widespread implementation of incentive regulation in the European electricity distribution sector (see, e.g., Jamasb and Pollit, 2001; Haney and Pollitt, 2009, 2011). In this sector the firms are natural monopolies, and their pricing policies are usually regulated by some government agency. The traditional cost-of-service or rate-of-return regulation is known to provide insufficient incentives for distribution system operators (DSOs hereafter) for cost efficiency. A number of European regulators have introduced benchmarking approaches such as data envelopment analysis (DEA) or stochastic frontier analysis (SFA) in order to create incentives for cost efficient operation (see e.g. Jamasb and Pollit, 2007; Kopsakangas-Savolainen and Svento, 2008; Bogetoft and Otto, 2011). The emphasis on cost efficiency has however created adverse incentives for DSOs to decrease the quality of their services (Joskow, 2008). Recently considerable interest has been placed on studying how incentive regulation affects the quality related investments and the quality of service in network industries (e.g., Ai et al., 2004; Cambini and Rondi, 2010; Reichl et al., 2008). Empirical evidence suggests that incentive regulation focusing only on operational costs can reduce the quality of service unless regulation is amended with some quality incentives also (Hafner et al., 2010; Ter-martirosyan and Kwoka, 2010). Thus it seems clear that the regulatory models must be complemented with quality regulation in order to maintain an acceptable level of supply security (see e.g. Jamasb and Pollit, 2008).

The quality of service is seen an important objective by the customers, industry and the regulator alike. Poor service quality such as supply interruptions often leads to losses for industry and households in terms of lost production or the lost utility that customers can obtain from the energy services (de Nooij et al., 2007). As the task of the government (regulator) is to guarantee stable conditions to operate for industry and households, the service quality is a concern for the regulator also.[1] Consequently it needs to be examined how firms can improve their quality of service. Investments on network are one of the more direct ways to affect security of supply as older equipment is replaced with newer one. The most pronounced investment type on how firms can affect their quality of service is underground cabling. For example Fenrick and Getachew (2012) identify underground cabling as a highly important factor in reducing interruptions. Less emphasis however has been placed on how underground cabling affects the variability of interruptions. Since customers (and regulator) can be viewed to be risk averse, they view not only the small level but also

* Corresponding author. Tel.: +358 9 43131; fax: +358 9 43138535.
E-mail address: antti.saastamoinen@aalto.fi (A. Saastamoinen).

[1] Customers' valuation of the interruptions of course partly depends from the customer type. See for example Sullivan et al. (1996) for an early discussion and de Nooij et al. (2007) for more recent study.

http://dx.doi.org/10.1016/j.eneco.2014.04.016
0140-9883/

the low variability of interruptions as a sign of good quality. Given a certain expected level of interruptions, the scenario with less variability would be favored by most customers over a scenario with high fluctuations in the duration and the frequency of interruptions as the former scenario would guarantee a more stable planning horizon. Risk aversion could be argued to be especially high in countries with highly variable weather conditions, such as Finland. Thus quality regulation should aim to reduce also the risk of interruptions in order to meet the customers' expectations of low variability. However, as Fenrick and Getachew (2012) state, the decision to invest on underground cabling is not straightforward as these investments incur extra costs compared to over-headlines. These costs include for example higher installment costs, costs due to longer repair times, and higher material costs (Hall, 2013). Thus the managers have to weigh the benefits of underground cabling against its extra costs. If the managers perceive the cost to be greater than the benefit, the level of quality may not be at the socially optimal level as managers probably do not consider the consumers' valuation of supply security when making investment decisions. There is large body of literature that discusses about the optimal *level* of quality in electricity distribution sector (Ajodhia and Hakvoort, 2005; Jamasb et al., 2012; Sappington, 2005). The variability of quality is however often neglected from these discussions. This study aims to shed light on how this variability can be affected by underground cabling investments. Our results suggest that underground cabling does not have a significant decreasing effect on the variability of interruption costs. In fact, the effect may be even risk increasing. From policy perspective this implies that firms may need to be given further incentives to undertake underground cabling investments.

Another issue is the practical implementation of quality regulation. Setting the target quality level is one important part of the implementation. In general regulation is challenging as firms usually have an informational advantage over the regulator about their true costs (see Holt, 2005; Kopsakangas-Savolainen and Svento, 2010; Sappington, 2005). Similarly to Shleifer's (1985) classic yardstick model of regulation, already Alexander (1996) discussed using the performance of comparable firms as a way to set the targets. However it may be difficult to find such comparable firms (Pollitt, 2005). Benchmarking methods are considered to overcome the problem of asymmetric information and finding an objective comparison point (see e.g. Ajodhia and Hakvoort, 2005). These methods however have not been used in the regulation of service quality as extensively as in the regulation of costs. For example in Finland the quality targets are set by averaging the own previous performance of the companies. Thus, if a DSO currently operates at a low quality level, it only needs to maintain its current low quality level without any need to improve its performance over time.

In this study we suggest that the best practice benchmarking methods could be utilized in setting the quality targets. We argue that the best practice is preferred to the average level, as the latter approach can create undesired incentives (see Ajodhia and Hakvoort, 2005). The industry wide performance is also likely to be improved more by using the best practices. We introduce a best practice method to be used in setting the quality target and compare it to the current practice of Finnish regulator. Our results indicate that the targets produced by the proposed method are more stable for DSOs of similar sizes than the targets obtained with the current approach of Finnish regulator. These findings seem to be in line with the DSO hopes of developing the foreseeability and stability of the regulatory model and improving the incentives for better performance found by Tahvanainen et al. (2012) in their survey (see also Kinnunen, 2006).

Methodologically both of the above aims, the examination of underground cabling effects and setting the quality targets, can be met by utilizing a recently developed StoNED method for frontier estimation (Johnson and Kuosmanen, 2011; Kuosmanen, 2012; Kuosmanen and Kortelainen, 2012). This estimation method non-parametrically estimates a frontier of quality performance what we call as a *quality frontier*. It also readily incorporates the effects of operational environment of DSOs into its estimation framework. It is generally well acknowledged that the operational environment of DSOs should be taken account in a typical benchmarking process. Network operators are subject to different weather conditions, geographical conditions, and consumer densities which affect their costs and service quality (see e.g. Growitsch et al., 2009, 2012; Simab and Haghifam, 2012; Yu et al., 2009a). In this work we consider the amount of underground cabling as measuring these operational conditions (see e.g. Kuosmanen et al., 2013; Kuosmanen, 2012). DSOs operating in a dense city areas have different underground cabling levels than DSOs in the rural areas. Thus the quality frontier presented in this study accounts for these differences in determining the proper quality targets.[2]

This paper is organized as follows. Section 2 briefly discusses the measurement of service quality and describes the theoretical quality frontier model framework and the empirical estimation method associated with it. Section 3 summarizes the data. In Section 4 we examine the effects of underground cabling on the level and the variance of interruption costs. Section 5 moves to examine the practical implications of using the estimated quality frontier instead of the current Finnish practice in quality target setting. This section also briefly describes the overall Finnish regulatory system. Section 6 then concludes.

2. Quality frontier model

This section introduces the quality frontier model and the necessary terminology and notation. The purpose of this section is also to address the questions of why a frontier model of quality is interesting and what type of information it can provide for the regulators. We also briefly discuss about the measurement of quality at this junction.

2.1. The measurement of quality

In this study, we use the costs of interruptions as the quality indicator (see e.g. Ajodhia, 2010; Growitsch et al., 2010). In Finland the interruption costs are calculated by the Finnish Energy Market Authority (*Energiamarkkinavirasto* (EMV)). The calculation takes into account the duration and number of interruptions. Thus in this study we are only concerned about the continuity of supply aspect of quality. Consequently we do not consider for example commercial or technical aspects of service quality, such as the quality of billing services and voltage variations. The estimates of customers' willingness-to-pay (WTP) to avoid interruptions or the valuation of lost energy are then used to transform the technical measures into costs (see e.g. Reichl et al., 2013; McNair et al., 2011; Growitsch et al., 2010; Yu et al., 2009b; de Nooij et al., 2007).[3] In Finland the customer valuation is based on the survey made by Silvast et al. (2005). The formula on how interruption costs are calculated by the Finnish regulator can be found from EMV (Finnish Energy Market Authority) (2011a) and from Appendix A of this study.

Alternative approach would be just to use technical measures common in the literature such as frequency and duration of outages, customer minutes lost or the loss of energy delivered (see e.g. Fernandes et al., 2012; Simab and Haghifam, 2012). Such technical measures can

[2] In Norway, a large set of environmental and operational condition variables are used in a traditional regression model to estimate an expected level of interruption cost which is then used as a reference value (see Langset et al., 2001). Kopsakangas-Savolainen and Svento (2011) consider load factor variable as a variable describing the operational environment.

[3] Alternative to WTP is willingness-to-accept (WTA), that is, how much customer should be compensated in order to accept an interruption of a certain size. Generally there is large disparity between WTP and WTA measures as the latter is often measured to be much larger than the former. WTA is heavily driven up by the loss aversion of the customers (see e.g. Beenstock et al., 1998). WTP on the other can be subject to underestimation (see e.g. Linares and Rey, 2013).

Please cite this article as: Saastamoinen, A., Kuosmanen, T., Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland, Energy Econ. (2014), http://dx.doi.org/10.1016/j.eneco.2014.04.016

be directly incorporated for example to a DEA model as one of the input variables (Giannakis et al., 2005; Yu et al., 2009a, 2009b). The likely problem of such approaches is that the firms may specialize to a certain type of quality or specialize either on quality or operational costs (see e.g. Ajodhia, 2006). That is, firms may seem efficient by only focusing either on cost reductions or quality improvement, but not necessarily on both. Some recent studies propose to combine multiple quality dimensions to a single quality indicator using DEA (Ferrier and Trivitt, 2012 (health sector); Simab and Haghifam, 2012 (electricity); Façanha and Resende, 2004 (telecommunications)). By forming a single quality index these studies attempt to take account the multi-dimensional nature of quality (see e.g. Fumagalli and Lo Schiavo, 2009). Such indices however might hide some specific aspects of quality and it could be challenging to distinguish from these indices that in what way quality should be improved. An economic measure of quality, such as interruption costs used in this study, accounts for both the technical aspects of and the customer valuation in forming the measure. In terms of social welfare, the economic approach makes it more straightforward to analyze whether the quality provision is at the socially optimal level. Of course, if the interest is to examine on which specific (technical) aspects of the quality the improvements should be targeted, then we should use a disaggregated analysis where each (relevant) quality component is separately included into the model. But since our aim here is to estimate the reference level within the Finnish system, we remain using the interruption costs as our measure of quality.

2.2. Theoretical quality frontier model

The conceptual framework of the quality frontier model is given Eq. (1). The total external supply interruption costs are given by the variable x. We assume that interruption costs depend on the outputs $\mathbf{y}$ and contextual variables $\mathbf{z}$ that characterize operational conditions and practices of distribution networks. For the sake of generality, we abstract from the definition of the output vector $\mathbf{y}$ and the contextual variables $\mathbf{z}$ and leave it for the regulators. Using these notations, the general model of quality frontier can be represented as

$$x = Q(\mathbf{y}) \cdot u(\mathbf{z}) \cdot v(\mathbf{z}) \tag{1}$$

where

x	is the interruption cost
Q	is the quality frontier function
$\mathbf{y}$	is the vector of outputs
$\mathbf{z}$	is the vector of contextual variables
u	is a random variable representing inefficiency
v	is a random variable representing stochastic noise.

The quality frontier function Q is directly analogous to the frontier cost function: it represents the minimum interruption cost at the given output level. We assume that the quality frontier Q is a monotonic increasing and convex function of outputs and that Q exhibits constant returns to scale (see, e.g., Kuosmanen, 2012, for further discussion of the axioms in the context of energy regulation). However, we do not impose any particular functional form for the quality frontier. For example the often used Cobb–Douglas form implies economies of specialization, which is problematic when modeling joint production. Electricity distribution companies are usually modeled as multi-output producers as often variables such as number of customers and network length are considered their outputs alongside the distributed electricity. On the other hand, using more flexible functional forms often violate for example monotonicity.

An important point to note in model (1) is that we assume the random inefficiency term u and the noise term v depend on the contextual variables $\mathbf{z}$. More specifically, we assume that the asymmetric inefficiency term u has the half-normal distribution that depends on the contextual variables $\mathbf{z}$ according to

$$u(\mathbf{z}) \sim N^{+}(0, \sigma_u(\mathbf{z})), \tag{2}$$

$$\sigma_u(\mathbf{z}) = \exp(\mathbf{z}'\boldsymbol{\theta}). \tag{3}$$

Note that the expected value of inefficiency depends on $\mathbf{z}$ according to

$$E(u(\mathbf{z})) = \mu(\mathbf{z}) = \exp(\mathbf{z}'\boldsymbol{\theta})\sqrt{2/\pi}. \tag{4}$$

Similarly, we assume that the variance of the noise term depends on contextual variables

$$v(\mathbf{z}) \sim N(0, \sigma_v(\mathbf{z})), \tag{5}$$

$$\sigma_v(\mathbf{z}) = \exp(\mathbf{z}'\boldsymbol{\gamma}). \tag{6}$$

In the terminology of econometrics, we assume that both the inefficiency and noise term are heteroscedastic. The stochastic part of the quality frontier model can be interpreted as doubly heteroscedastic model introduced by Hadri (1999, 2003) (see also the recent survey of heteroscedastic SFA models by Alvarez et al., 2006).[4] Why should one care about heteroscedasticity in the present context? There are at least two good reasons why the regulators and the regulated firms should care.

Firstly, as already stated in Section 1, the operational conditions and practices can affect the risk of interruptions, commonly measured by the variance. For example, the use of underground cables instead of overhead cables can make the network less vulnerable to storms and other extreme weather events. Note that customers of electricity distribution networks are typically more risk averse than the firms providing the service. Risk neutral firms may be willing to tolerate higher risks than their risk averse customers, leading to a suboptimal investment to underground cabling if the risk effect is ignored. If the elements of vector $\mathbf{z}$ are controlled by firms, then the quality frontier model can help the regulators to create better incentives for improving the quality of service through the $\mathbf{z}$-variables.

Secondly, even if one is only interested in the expected value of interruption cost (e.g., all relevant parties are risk neutral) and even if some (or all) elements of $\mathbf{z}$ are considered uncontrollable, it is important to take the variance into account from the econometric point of view. This is because the shape of the quality frontier Q will generally differ from that of the conditional mean function $E(x|\mathbf{y},\mathbf{z})$ if the inefficiency is heteroscedastic. Therefore, the usual methods of regression analysis provide biased estimates if the heteroscedasticity effect is ignored (see Florens and Simar, 2005, for further discussion).

The following simulated example illustrates the second point. In this example we assume the true quality frontier as $Q(\mathbf{y}) = \mathbf{y}^2$. We assume uniformly distributed $\mathbf{y}$, and a single contextual variable that is uniformly distributed with $Cov(\mathbf{z},\mathbf{y}) = 0.9$. The standard deviation of the half-normally distributed inefficiency term u is $0.4\mathbf{z}$ and that of the normally distributed noise term v is $0.2\mathbf{z}$. We draw a random sample ($n = 200$), and add inefficiency and noise to the quality frontier $Q(\mathbf{y})$. The true frontier Q (the black curve) and the scatter of the simulated data points are presented in Fig. 1. The OLS estimate of the quadratic function is presented in the figure by a gray broken curve (the estimated equation is $x = 1.26\mathbf{y}^2 + 0.35\mathbf{y} - 0.23$; $R^2 = 0.93$). The figure aptly illustrates

[4] The model by Wang (2002) parameterizes both the mean and the variance of the inefficiency distribution with $\mathbf{z}$-variables. His model accommodates non-monotonic efficiency effects i.e. $\mathbf{z}$-variables can have different effects at the different levels.

Please cite this article as: Saastamoinen, A., Kuosmanen, T., Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland, Energy Econ. (2014), http://dx.doi.org/10.1016/j.eneco.2014.04.016

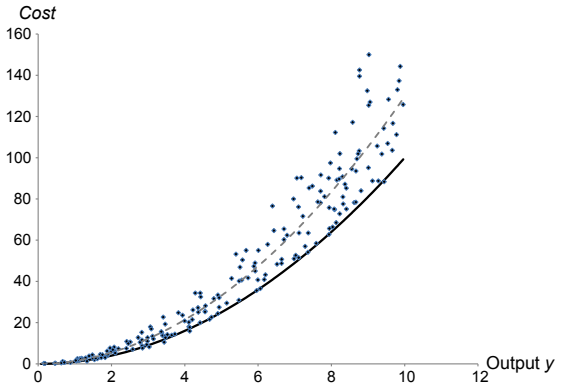

Fig. 1. Illustration of the heteroscedasticity effect. Solid black curve is the true frontier $Q = \mathbf{y}^2$ used in the simulation. The gray broken curve is the OLS estimate. The shapes of the two curves differ due to heteroscedasticity.

that the shape of the OLS curve differs notably from that of the true quality frontier *Q*.

2.3. Semi-nonparametric estimation[5]

To estimate the quality frontier model (1) in a semi-nonparametric fashion without making additional assumptions to those stated in the previous section, we resort to the StoNED approach (Kuosmanen, 2012; Kuosmanen and Kortelainen, 2012). As a starting point, we take the logarithms of both sides of Eq. (1) and rewrite it as the partially linear model

$$\ln x = \ln Q(\mathbf{y}) + \mathbf{z}'\boldsymbol{\delta} + \varepsilon(\mathbf{z}), \tag{7}$$

where $\mathbf{z}'\boldsymbol{\delta} = E(\ln u(\mathbf{z}))$ and the composite error term

$$\varepsilon(\mathbf{z}) = \ln u(\mathbf{z}) + \ln v(\mathbf{z}) - \mathbf{z}'\boldsymbol{\delta} \tag{8}$$

has zero mean. Therefore, the quality frontier *Q* and the effects of z-variables on the expected value of *x* can be consistently estimated by the semi-parametric CNLS estimator (Johnson and Kuosmanen, 2011, 2012). Note that the coefficients $\boldsymbol{\delta}$ can be interpreted as the post-truncation effects of the **z**-variables on the expected inefficiency, whereas coefficients $\boldsymbol{\theta}$ introduced in Eq. (3) represent the pre-truncation effects. In other words, $\boldsymbol{\delta}$ and $\boldsymbol{\theta}$ are just alternative parameterizations of the same effects.

In stage 1, we solve the following nonlinear programming problem shown in Eq. (9):

$$\min_{\phi,\beta,\delta,\varepsilon} \sum_{i=1}^{n} \varepsilon_i^2 \tag{9}$$

s.t.

$$\ln x_i = \ln \phi_i + \sum\nolimits_k z_{ik}\delta_k + \varepsilon_i \quad \forall i$$
$$\phi_i = \beta_{1i}y_{1i} + \beta_{2i}y_{2i} + \beta_{3i}y_{3i} \quad \forall i$$
$$\phi_i \geq \beta_{1h}y_{1i} + \beta_{2h}y_{2i} + \beta_{3h}y_{3i} \quad \forall h, i$$
$$\beta_{ki} \geq 0 \forall k = 1,2,3; \forall i.$$

In the set of Eq. (9), all the variables are defined as earlier. The parameter ϕ_i is the frontier interruption cost for firm *i*. Note that we include time dummies into the set of contextual variables. The first constraint defines the regression equation. The second set of constraints specifies that the tangent hyperplanes of the frontier are linear. These constraints do not however restrict the form of the frontier in any way. The third and the fourth sets of constraints define the cost function to be convex and monotonically increasing in outputs. The resulting frontier is piece-wise linear and it is very flexible in terms of technology as the marginal costs of outputs (β_i) are firm specific.

Given the CNLS residuals $\hat{\varepsilon}_i$ from the problem in Eq. (9), we can apply the quasi-likelihood approach by Fan et al. (1996) to estimate the doubly heteroscedastic inefficiency model by Hadri (1999). In stage 2, we solve the quasi-likelihood problem, for which, following Hadri (1999), the log-likelihood in terms of the CNLS residuals can be written as in Eq. (10). It is assumed here that inefficiency and noise are distributed according to Eqs. (2) and (5).

$$\log L(\beta, \delta, \boldsymbol{\gamma}) = \sum \log(f_i(\hat{\varepsilon}_i)) \tag{10}$$

where

$$f_i(\hat{\varepsilon}_i) = (2/\sigma_i) f^*(\hat{\varepsilon}_i/\sigma_i) F^*(\lambda_i \hat{\varepsilon}_i/\sigma_i)$$

$$\sigma_i = \sigma_{vi} + \sigma_{ui}$$
$$\lambda_i = \sigma_{ui}/\sigma_{vi}$$

where f^* and F^* are the standard normal density and distribution functions. This problem can be computed with any standard software packages (for example, Stata), which allows the parameterization of both standard deviations (variances) separately in a frontier model. In practice, we can estimate the parameters of the second stage inefficiency model by regressing the equation

$$\ln x_i - \ln \hat{\phi}_i = \alpha + \sum\nolimits_k z_{ik}\tilde{\delta}_k + \tilde{\varepsilon}_i \tag{11}$$

applying standard computational tools for SFA models. Note that values $\hat{\phi}_i$ on the left-hand side of Eq. (11) are obtained from the optimal solution to Eq. (9), and hence the left-hand side of Eq. (11) is given beforehand. The parametric inefficiency model in the second stage serves to identify the effects on different variance components, as it is computationally prohibitive to incorporate the variance effects in the mathematical programming problem (Eq. (9)). Furthermore the second stage provides us with the typical standard errors to access the statistical significance of the effects. These are not readily available in the first nonparametric stage. Notice that if your interest is only on the significance of the level effects (parameters $\tilde{\delta}_k$), we could estimate Eq. (11) by OLS and adjusting the standard errors for possible heteroscedasticity. As a consequence we estimate the level effects in both ways as this serves as an internal consistency check of our results. The estimates of $\tilde{\delta}_k$ parameters from Eq. (11) should be relatively the same regardless of whether linear regression or the parametric inefficiency model is applied in the second stage. They will not be exactly the same as the inefficiency model includes a further parameterization of the variances.

Lastly, we stress that the estimation framework followed here is not a typical two-step method that has been heavily criticized in the literature (see, e.g., Wang and Schmidt, 2002). The critique concerns such two-stage methods where the **z**-variables are neglected altogether in the first stage estimation, creating a possible omitted variable problem. Note that in our approach the effects of the **z**-variables on the conditional mean of the dependent variable are duly taken into account in the first stage estimation. Moreover, the benefit of our methods is that we estimate the effect of operational conditions and the quality frontier at once. The standard two-stage methods are suitable for estimating the effects of **z**-variables. However, some further stages would be required

[5] See Kuosmanen et al. (2014) for a detailed treatment of the estimation framework. They also deal on some ways to model heteroscedasticity within the framework.

Please cite this article as: Saastamoinen, A., Kuosmanen, T., Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland, Energy Econ. (2014), http://dx.doi.org/10.1016/j.eneco.2014.04.016

Table 1
Summary statistics of the variables used (period 2005–2010; n = 516).

Variable	Mean	Std. dev.	1st quartile	Median	3rd quartile	Min	Max
Energy (GWh)	503.33	1021.01	73.99	173.47	446.7	14.54	7297.84
Network (km)	4297.62	10425.78	720.05	1055.00	3431.1	26.30	70096.90
Number of users (1000 s)	37.08	73.62	4.98	12.12	27.61	0.02	441.49
Underground cabling (%)	22.56	27.58	2.70	7.95	38.5	0.01	100.00
Interruption cost (mill. €)	1.46	4.34	0.11	0.30	0.82	0.0007	44.90

in order to take these effects into account in estimating the quality targets via frontier (see e.g. Fried et al., 1999, 2002). Such multi-stage procedure may not be desirable in terms of transparency of the regulation. Further, the first stage estimator remains unbiased and consistent even if there is heteroscedasticity with respect to the **z**-variables. Therefore, it is perfectly valid to reduce the heavy computational burden of the first stage nonparametric estimator by leaving the variance effects to the second stage regression.

3. Description of the application and data used

The data have been obtained from the website of the Finnish regulator (EMV).[6] The data is a balanced panel consisting observations of 86 DSOs over the period 2005–2010, making it total of 516 observations. We have data on interruption costs, underground cabling and three outputs, namely energy transmission, network length, and number of customers. These outputs are commonly considered as the primary cost drivers for DSOs (see e.g. Korhonen and Syrjänen, 2003; Giannakis et al., 2005; Thakur et al., 2006). We consider the same outputs to be the drivers of interruption costs also. In other words we argue that the scale and scope of operations should dictate the reference level of interruption costs. The summary statistics of the variables are given in Table 1.

The energy transmission output is a weighted sum of transmitted energy at every voltage level. The weights are defined as the average cost of energy transmission and the transmission at the lower voltage level gets a higher weight such that the transmission is measured in GWh of 0.4 kV equivalents. That is, the 0.4 kV transmission gets a weight of one and higher voltage levels weigh below one. Underground cabling is the proportion of underground cabling in 1–70 kV network. This variable is used by EMV as a contextual variable to describe the operational conditions of DSOs. In addition to the variables in Table 1 we have the geographical location of DSO as an additional **z**-variable. This has been approximated with the coordinates of DSOs' head office.[7] The information is obtained from EMV as DSOs are required to report their contact details. This proxy is rather rough approximation of the location as some DSOs operate on very large areas or even on multiple areas which are geographically separate. Typically operators however have their offices within the same area that they operate in. For the majority of companies in our data, this proxy of their location is straightforward to define. There are two companies which have two (or more) distinctively separate geographical operation areas. For these two companies the location has designated so that if the undergrounding level of the company is low, the company has been designated a more rural coordinates and vice versa. Of course, it would be a concern if the data would include many such companies with operating areas separated by long distances. In this case, identifying the effect of location with coordinates would be difficult. Below we however will see that our assumptions about the locations of the abovementioned two companies are rather unimportant (see footnote 7). Alternatively we could have used some arbitrary division of the country for example into four regions in terms of cardinal points. But such division would assign many companies the same location even though the area might include for example coastal companies and inland companies.

Table 1 illustrates the structure of electricity distribution industry in Finland. Few larger companies with significantly larger outputs cause the output distributions to be heavily skewed to the right (skewness statistic not shown, but it was positive for all variables). For example, for 75% of observations energy transmitted is less or equal to 446.65 GWh. The largest company has transmitted approximately 7300 GWh per year. The interruption costs vary from minimal €714 to almost €4.5 million within the period 2005–2010. The yearly summary statistics in online Appendix A show that 2010 was a stormy year with high interruption costs.[8] Importantly, the high interruption costs observed in 2010 will inflate the EMV reference values as average is sensitive to outliers. This leads to more lenient targets. If the year for which the target is calculated is also stormy, such target would be adequate. But if interruptions revert back to their "normal" level after 2010, the target would be overstated. Lastly note that our data includes few industrial networks. They have rather low number of customers and short network length but relatively large energy transmission.

Since the relationship between interruption costs and underground cabling is our main target of interest, we illustrate their connection in our data with Fig. 2. The observed interruption costs for each year (2005–2010) have been plotted against the underground cabling levels. As expected, the largest variability in interruption costs is at the low cabling level. These are the companies with large overhead networks. There is a slight downward trend in interruptions as the cabling level increases, implying that the level of interruptions decreases along with underground cabling. Nevertheless there are relatively large variations in the observed interruption costs at the higher levels of cabling also. For example at the level 50% of underground cabling, the log of interruption costs might vary from 10 to 15 (from €22,000 to almost €3.3 million in actual monetary terms). We also see some observable variation at the very high levels of cabling. The major part of the variation at these high levels is due to observations with 100% cabling. These six observations belong to one of the abovementioned industrial network which has 100% cabling proportion. We also conduct a robustness check of our results when we exclude this (and one other) industrial network from the sample (see online Appendix C).[9] The analysis in the next section includes these two firms.

4. Estimated effects of underground cabling

We first study how underground cabling affects the level of interruption costs in Table 2. We present the results for two different model specifications. Model 1 includes the underground cabling and the year dummies as **z**-variables. Model 2 includes the coordinates of DSOs' head office (i.e., latitude, longitude) in addition to the underground cabling and year dummies. The alternative model specification tests the robustness of the results and whether there is any other location specific effects that underground cabling does not identify.

We present the direct estimates of the level effects from the StoNED model and the estimates of the second stage parametric inefficiency

[6] The webpage of EMV is: http://www.emvi.fi/.

[7] Coordinates have been obtained from Google Maps based on the city/town that the company has the head office.

[8] See also EMV (Finnish Energy Market Authority) (2011b).

[9] The network operators that have been removed from the results of online Appendix C are an operator that provides services only for an airport and one which serves only an industrial park.

Please cite this article as: Saastamoinen, A., Kuosmanen, T., Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland, Energy Econ. (2014), http://dx.doi.org/10.1016/j.eneco.2014.04.016

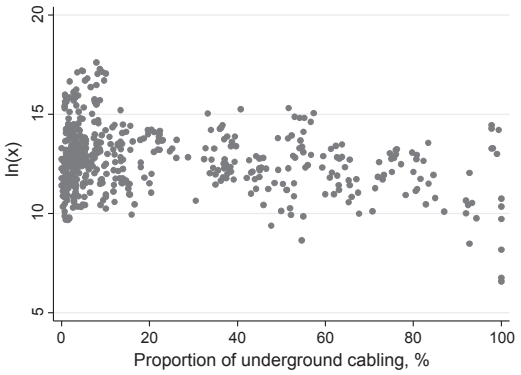

Fig. 2. Scatterplot, ln(x) vs. underground cabling (x = observed interruption costs).

model (Eq. (11)) (see discussion in Section 2). The standard errors in StoNED results are adjusted for heteroscedasticity since we expect interruption costs to be heteroscedastic with respect to underground cabling. In the second stage parametric inefficiency model heteroscedasticity is explicitly taken into account by parameterizing the error component variances.

Overall the results are rather stable over all specifications. Underground cabling significantly decreases the level of interruption costs in all models. Except the *longitude* variable in the StoNED Model 2 the coordinate variables are mainly insignificant in explaining the level of interruption costs. Underground cabling is already likely to capture the most characteristics of the operating environment. Moreover, part of the heterogeneity is modeled through variances of the error components in the second stage inefficiency model.

The *longitude* variable in StoNED, Model 2, has the expected sign. The Eastern Finland is likely to have higher interruption costs as it has more forests and overhead cables. Subsequently it is more likely to observe interruptions by trees falling on cables.[10] Notice that when we take the variance effect into account as in Model 2 of the 2nd stage inefficiency model, this effect becomes insignificant. The yearly dummies indicate a technical progress in the industry as their coefficients are negative. In fact the industry level underground cabling proportion has slightly increased over the period. Hence it is expected that the interruption costs have declined. The coefficient for the year 2010 is surprising. Knowing that 2010 was an exceptional storm year we would expect a positive sign and probably a highly significant effect. We suspect that the yearly dummies capture more the overall trend than any exceptional events. Partly the exceptional storms of 2010 are manifested in the coefficient as it is not large compared to the other yearly dummies. We interpret that the insignificance of this coefficient implies that the effect of improved technology has been insignificant because of the extreme weather conditions in 2010.

Next we examine the variance effects. For the sake of completeness and as an internal consistency check we also examine the variance effects with the linear regression. In that case we are restricted to study the effects on the estimates of the overall error obtained from solving problem (9). That is we do not yet separate the effects between noise and inefficiency variances. To examine the overall effects we in practice regress the squared estimates of the overall error on the contextual variables. This is the standard practice in econometrics to study heteroscedasticity (see e.g. White, 1980; Greene, 2008). The results of this analysis are presented in Table 3. In Table 4 on the other hand we explicitly differentiate the variance effects between inefficiency and noise in the second stage parametric inefficiency model and examine the parameter estimates of functions (3) and (6). Notice that the year dummies have been excluded from these models.[11] We stress that the parameter estimates in Tables 3 and 4 are not directly comparable because of the differences in the variance parameterization.

Table 2
Level effects on the level of interruption costs.

	StoNED		2nd stage inefficiency	
z-Variable	Model 1	Model 2	Model 1	Model 2
U. cabling	−0.025***	−0.026***	−0.019***	−0.022***
	(0.002)	(0.002)	(0.003)	(0.003)
Latitude		−0.029		0.040
		(0.015)		(0.023)
Longitude		0.086***		0.018
		(0.015)		(0.024)
2006	−0.207**	−0.209**	−0.159	−0.201
	(0.096)	(0.094)	(0.132)	(0.118)
2007	−0.323***	−0.324***	−0.319***	−0.347***
	(0.102)	(0.100)	(0.119)	(0.112)
2008	−0.260**	−0.259**	−0.275**	−0.310***
	(0.107)	(0.105)	(0.115)	(0.110)
2009	−0.537***	−0.534***	−0.591***	−0.611***
	(0.107)	(0.104)	(0.115)	(0.109)
2010	−0.133	−0.137	−0.196	−0.172
	(0.121)	(0.117)	(0.107)	(0.103)
Intercept	0.623***	0.416	0.013	−2.557
	(0.074)	(0.969)	(0.117)	(1.373)

Standard errors in parenthesis.
*** 1% significance.
** 5% significance.

Table 3 reveals that the variance of interruption costs increases as the level of underground cabling increases. We would expect lower variability with higher proportion of underground cabling. But the companies which have a high proportion of underground cabling also have a large customer base and large energy transmission. Thus when these companies face an interruption, their costs are likely to be much higher than their usual levels since a great number of customers are affected by the interruption. Consequently the variance of interruption costs can be high for these companies as we only observe small and some high interruption costs. In addition the higher initial installment costs and the repair costs of underground cabling may be translated into interruption costs. The maintenance costs of underground cables are generally lower than the corresponding costs of overhead lines.

In Table 4 the variance of interruption cost has now been decomposed into two parts. Underground cabling has remarkably different effects on the variability of inefficiency and noise. Whereas cabling (insignificantly) decreases the variability of inefficiency, it significantly increases the variation in noise. Beforehand there is no reason to assume any specific sign on the first effect. The negative sign indicates that variation of (in)efficiencies among high underground cabling companies is smaller. We could argue that the companies using mainly overhead lines experience higher variations in their daily operations and consequently they might have higher variations in their efficiency. The positive effect of underground cabling on noise on the other hand

[10] Notice that the effect of coordinate variables could be partially driven by our assumptions of the locations of two companies with rather high interruption costs. Thus we conducted an OLS analysis, where we regressed interruption costs on underground cabling and the coordinates. We varied the values of coordinates in terms of different locations of these companies in order to examine whether this had any major effect on the coefficients of the coordinate variables. None of the estimates experienced major changes in sign or significance. Thus we conclude that these location assumptions do not affect the overall results stated above. The results of these estimations can be obtained from the authors by request.

[11] This reduces the number of estimable parameters in maximum likelihood. In fact, in the StoNED models, if we include year dummies in the models of Table 3, the coefficient for 2010 dummy was positive and significant. This is expected as 2010 was a storm year and the variability of interruption costs was expected to be high.

Please cite this article as: Saastamoinen, A., Kuosmanen, T., Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland, Energy Econ. (2014), http://dx.doi.org/10.1016/j.eneco.2014.04.016

Table 3
The variance effects; overall error.

z-Variable	Model 1	Model 2
U. cabling	0.007***	0.007***
	(0.002)	(0.002)
Latitude		0.018
		(0.024)
Longitude		0.053**
		(0.025)
Intercept	0.401***	−2.083
	(0.071)	(1.521)

Standard errors in parenthesis.
*** 1% significance.
** 5% significance.

implies higher variability of interruption costs due to random phenomena such as weather. Therefore the effect of weather on variability seems to increase along with cabling. For example Growitsch et al. (2012) found that the higher amount of energy delivered strengthens the cost increasing effect of poor weather. Recall that the companies with a higher underground cabling proportion often deliver more energy than the companies with overhead lines. Comparing Tables 3 and 4 we notice that the effect on noise variability drives the effects on the overall error since only noise effect is significant in Table 4. This finding confirms the internal consistency of our estimation framework as the effect on overall error is positive. The results after removing two industrial networks considered as outliers from the data are given in online appendix. These results can be briefly summarized here. The level effects were very robust and they stayed negative and significant. The variance effects turned to be insignificant. Notably however we did not find any significant negative effect either. That is, in our data, the effects of underground cabling on the variance of interruption costs seem to be either positive or negligible, but not negative as we would expect beforehand.

From the practical policy perspective the results above show that it is worthwhile to promote underground cabling to further increase the security of electricity supply. Clearly the level of interruption costs can be decreased with investments on better cabling. Nevertheless regulator should also acknowledge that DSOs are likely to resists some of the demands to invest in underground cabling due to its costliness. More specifically, if the investments on underground cabling are in some sense at the saturated level already, the DSOs may consider further investments not worthwhile as they do not significantly improve their performance in terms of variability. Indeed our results suggest that underground cabling does not necessarily decrease the variability of interruption costs although the level is clearly negatively affected.

5. Finnish regulatory framework and reference value comparison

In this section we compare the reference values/quality targets obtained from the proposed quality frontier to the target values which are obtained using the current practice of Finland. We first briefly outline the Finnish regulatory framework and its quality incentive component. This illustrates the positioning of the quality targets within the system. The comparison of the target values is conducted with summary statistics and graphical illustrations.

5.1. Finnish regulatory model

We keep the description of the regulatory model relatively brief as detailed description of the system can be found for example from the document of EMV (Finnish Energy Market Authority) (2011a) and also from Tahvanainen et al. (2012).

The regulation of Finnish electricity DSOs is based on the rate of return regulation. In practice EMV determines the allowable returns for companies. The allowable returns are compared to the actual realized returns of the companies by taking the difference between them. If excess returns (positive difference) are observed, the DSO is expected to compensate the excess return by cutting the distribution prices in the next regulatory period. In contrast, if the actual returns are smaller than the allowable returns, the DSO has room for price increases. The cost efficiency and quality incentives adjust the observed return of DSOs. They are constructed such that it is beneficial for firms to meet (or pass) the targets set by these incentives. In the quality incentive component this means that it is beneficial for the company that the difference between the target value and the observed interruption cost is positive. In this way their difference is subtracted from the overall observed return. This again increases the likelihood that the observed return is lower than the regulated return. The same mechanism applies to the cost efficiency incentive. The simplified illustration of the system is given in Fig. 3.

Few comments are worth noting about the overall regulation framework. First, in the current regulatory period (2012–2015), the efficient cost frontier in the cost efficiency incentive is estimated with the similar StoNED model that this study proposes to be used also within the quality incentive. Secondly the interruption costs are divided between the cost efficiency and quality incentives. Inclusion of interruption costs into other cost (such as operational costs) is called the social total cost (SOTEX) approach in the literature (see e.g. Growitsch et al., 2010). It is aimed to mitigate the specialization problem discussed already in Section 2.1 as firms cannot neglect the quality considerations in the cost efficiency incentive component. One might also be concerned whether the overall cost level should be accounted for in determining the target interruption cost level. Obviously, in reality interruptions and their costs are dependent from the operational and capital costs of the companies. Indeed, it would be technically possible to consider a multi input multi output model where each cost type is separately included as an input. Alternatively other cost types could be accounted for by using the SOTEX approach outlined above. The main reason why we follow the suggested approach is to keep our approach as close as possible to the current regulatory model of Finland which includes an independent quality incentive component, which is our interest here. Note that the scale and scope of firms' operations are accounted for by the fact that interruption costs are determined as a function of outputs. Thus in this respect not accounting for other costs is not crucial for our approach.

The formula on how EMV currently calculates of the reference value is shown in Eq. (12). It is an energy transmission weighted average of past interruption costs. In the current period, the target for certain years is based on the values of 2005–2010. Notice also that the reference value as such does not yet represent the actual size of the whole incentive component. Before implementation the overall incentive component is subject to some further modifications, such as capping the maximum size of the penalty. The reference level however indicates how strict the quality regulation is when different reference levels are implemented. The higher the reference level is, the greater interruption costs are allowed.

$$IC_{ref,k} = \frac{\sum_{t=2005}^{2010} IC_{t,k} \times \left(\frac{W_k}{W_t}\right)}{6}, \tag{12}$$

where

$IC_{ref,k}$ is the reference level for year k
$IC_{t,k}$ is the IC for year t in the monetary value of year k
W_k is the transmitted energy in year k
W_t is the transmitted energy in year t.

Please cite this article as: Saastamoinen, A., Kuosmanen, T., Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland, Energy Econ. (2014), http://dx.doi.org/10.1016/j.eneco.2014.04.016

Table 4
Decomposition of variance effects in the second stage parametric inefficiency model.

	Model 1		Model 2	
z-Variable	θ inefficiency effects	γ noise effects	θ inefficiency effects	γ noise effects
U. cabling	−0.031	0.017***	−0.046	0.018***
	(0.025)	(0.002)	(0.038)	(0.003)
Latitude			−0.787**	0.133**
			(0.347)	(0.054)
Longitude			0.471***	0.001
			(0.163)	(0.058)
Intercept	−0.427**	−1.425***	35.690	−9.738***
	(0.207)	(0.153)	(19.097)	(2.885)

Standard errors in parenthesis.
*** 1% significance.
** 5% significance.

5.2. Comparison of quality targets

We start the comparison with summary statistics of the reference values in Table 5. The estimated reference values from the quality frontier for two different model specifications are identified as Est. IC ref (1) / (2). Note that the estimator discussed in Section 2 estimates a target value for each observation. Therefore in the analysis of this section we choose to use the targets for the observations that correspond to the year 2010. That is, we pick only one reference value for each firm. As a consequence we have 86 observations from where the summary statistics below are calculated.

On average the EMV method produces more lenient targets for companies. Like we already suggested in Section 3, some large interruption costs on 2010 for some companies have inflated EMV reference levels on average. The EMV method is more volatile as the standard deviation of EMV reference levels is substantially higher than the corresponding standard deviation of the quality frontier estimates. The estimated reference levels from Model 1 and Model 2 follow rather similar distributions. In the following analysis we only examine the values from Model 1 as results stay the same with Model 2.

Next we examine the stability of reference levels over the firms. Within regulation it can be argued that the reference level for companies of similar size should also be similar. In Fig. 4 we have plotted the log-transformed EMV reference values and the log-transformed reference values from the quality frontier. The observations have been ordered according to log-transformed average transmitted energy (2005–2010) in increasing order. Clearly the reference levels from the quality frontier are more stable between companies of similar size than the EMV reference levels. In log-terms the EMV reference values may vary from 10 to over 14 in small range of company sizes. In real monetary terms such differences mean a range of targets from €22,000 up to €1.2 million.

One can argue that the reference levels of two similar sized companies should differ as they might operate in a very different environment. Or vice versa, the reference level of similar sized companies operating in a similar environment should be relatively close to each other. In Fig. 5 we compare the reference levels separately at different levels of underground cabling. The firms have been grouped into four groups according to the average underground cabling proportion over 2005–2010 (online Appendix B describes how the groups have been formed). Once again the firms have been ordered according to their size in terms of transmitted energy.

Fig. 5 shows a clear difference in the variation of reference levels. The EMV reference values vary substantially more than the quality frontier levels even among companies of similar size and underground cabling level. Notably the quality frontier estimates of the reference levels for the high underground cabling group are substantially higher. We would except that these companies should have a rather strict reference

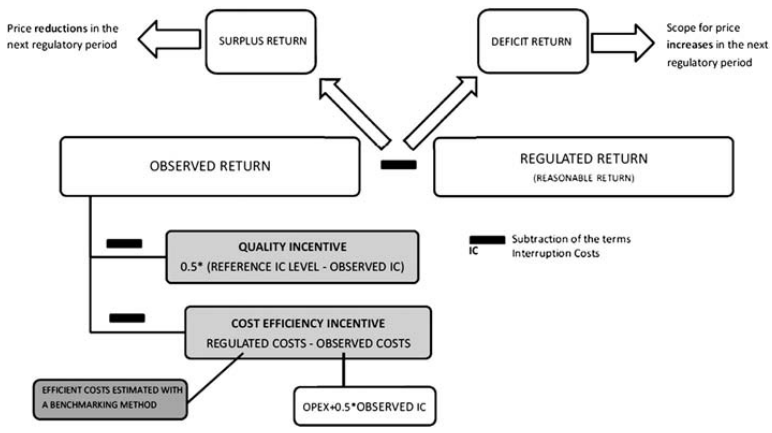

Fig. 3. The Finnish regulatory model.

Please cite this article as: Saastamoinen, A., Kuosmanen, T., Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland, Energy Econ. (2014), http://dx.doi.org/10.1016/j.eneco.2014.04.016

Table 5
Summary statistics of reference levels (n = 86), mill. €.

Reference value	Mean	Std. dev.	Skewness	1st quartile	Median	3rd quartile	Min	Max
Est. IC ref (1)	1.07	2.12	4.12	0.16	0.38	0.81	0.03	13.60
Est. IC ref (2)	0.97	1.92	4.18	0.15	0.34	0.76	0.03	12.50
EMV IC ref	1.47	4.04	4.86	0.14	0.28	1.01	0.02	25.80

levels since their usual level of interruptions is low. However since the typical level of interruptions is low for these companies, the EMV averaging approach might produce too strict targets for a year with exceptionally high interruptions. When a company with the high underground cabling level is actually hit by interruptions, the interruptions generally are large scale and hard to fix. This leads to substantial interruption costs.[12] This result illustrates that the proposed semi-nonparametric estimation of quality frontier better reflects the large scale and scope of operations of these companies in its determination of the target value. Finally, when we did your robustness check of the results, we found similar pattern of reference values between EMV values and the quality frontier values. The level of reference values was affected by the removal of the two industrial networks, but quality frontier still produced more stable targets than the EMV approach.

6. Conclusions

We have applied insights from productivity analysis to develop a frontier model of service quality. To estimate the quality frontier from interruption cost data, we proposed a new semi-nonparametric method, which does not require any functional form assumptions for the quality frontier. The method takes into account stochastic noise and heteroscedasticity effects both in the inefficiency and noise term. The proposed quality frontier was argued to provide more meaningful and stable basis for setting quality targets and incentives than the average practice benchmarks currently in use.

The empirical objectives of this study were twofold. Our first empirical aim was to study how the level of underground cabling and operational conditions affect the level and variance of interruption costs. As expected, underground cabling significantly explains the level of interruption costs. Interruption costs decrease with higher underground cabling levels. The effects on the variance of interruption costs are either positive or insignificantly different from zero. This implies that underground cabling does not significantly decrease the variability of interruption costs. This is because of the higher costs associated with the interruptions in underground networks. Even after robustness check we do not find a significant negative (risk decreasing) effect of underground cabling to the variability of interruption costs. Thus further investments in underground cabling might be perceived as unnecessary by DSOs after a desired level of interruption costs it attained. We also find that the variability of inefficiency is related to the geographical location of DSOs. This suggests that the performance differentials between DSOs are location specific. From the regulatory perspective this gives information for the regulator to characterize areas of relatively similar performance and areas of high variance of performance.

From the practical policy perspective, the variability of interruptions could play more significant role in the regulatory model in future. Similarly to the quality incentives based on the level of quality, we could amend the incentive system so that the low variability companies are rewarded whereas the high variability companies are penalized. This would be relevant especially in Finland, where high supply security requirements in future anyhow bring the overall level of interruptions down and more emphasis could be placed on the variability of interruptions as a regulatory channel.

Our second empirical objective was to compare the two alternative ways to set the quality improvement targets in the context of Finnish electricity distribution regulation regime. The current Finnish approach is based on averaging the previous performance of a DSO itself. This can be problematic for setting long term regulatory goals as poor previous performance might be translated to inadequate targets. Averaging is also susceptible of too high variation due to the influence of single years of high interruptions. Instead we suggest that target quality level should be set using a best practice benchmarking method. We estimate a quality frontier that can be interpreted to give the minimum interruption cost at the given output level. The estimated frontier produces more stable quality targets for similar sized companies than the current approach of Finnish Energy Market Authority. The quality frontier also explicitly accounts for the operating environment of companies.

For practical regulation in Finland, the use of quality frontier would make the quality regulation coherent with the cost efficiency regulation which is based on best practices. More importantly using quality frontier would make regulation more stable and equal. The overall best performance is likely to change less over time than the individual performance of a single company and thus the planning of supply performance would be easier. Since quality frontier produces similar targets for similar companies, the regulation can be considered to be more equal also.

As always, some limitations apply. First, we have used a rather limited set of contextual variables in our study. We have for example excluded customer density and weather variables from the study. Thus, no explicit conclusions about the effects of these factors should be made from our study. However, underground cabling and coordinate variables in practice already characterize these aspects of the operating environment. Underground cabling is likely to correlate strongly with population density. We considered only underground cabling variable as the regulatory model in Finland includes only this variable.

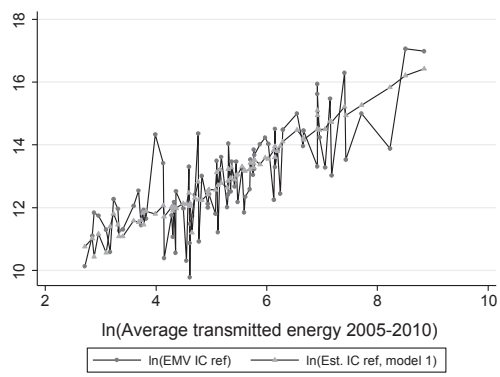

Fig. 4. The EMV reference levels against the quality frontier reference levels.

[12] Recall however that the effects of such severe conditions have been mitigated in the regulatory model by capping the maximum size of penalty from quality incentive component.

Please cite this article as: Saastamoinen, A., Kuosmanen, T., Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland, Energy Econ. (2014), http://dx.doi.org/10.1016/j.eneco.2014.04.016

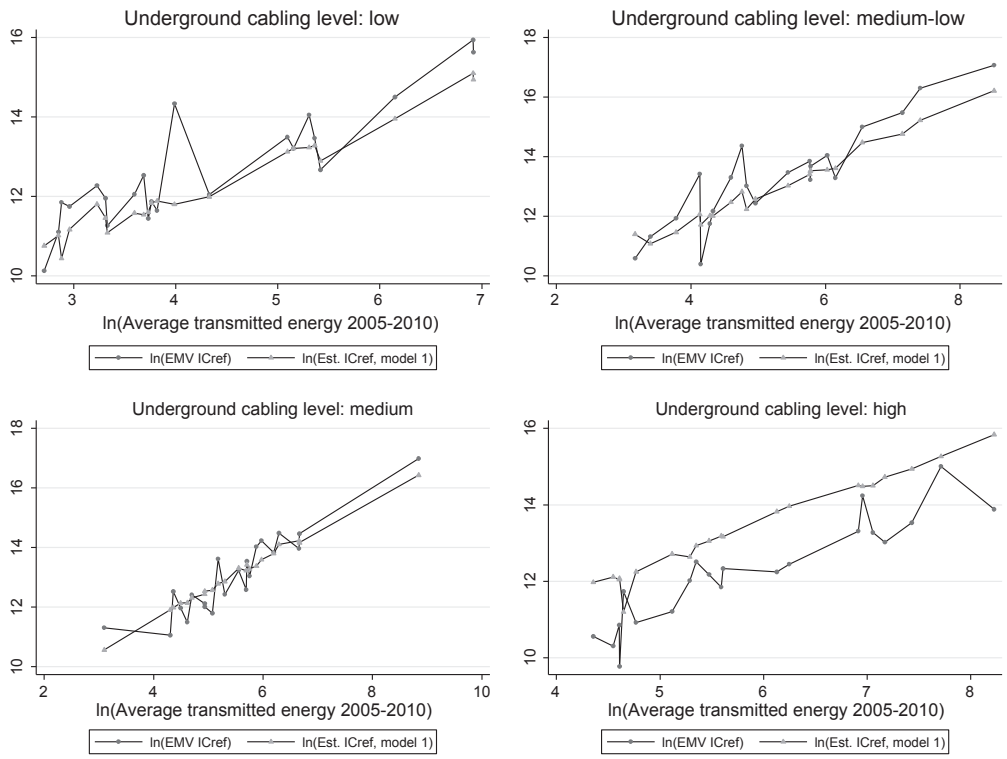

Fig. 5. The reference level comparison according the underground cabling level.

Second, we model the effects of **z**-variables as linear functions of underground cabling and the coordinates. Some nonlinear alternatives and interactions between the variables should be considered in future research in order to examine whether the effect of underground cabling differ between regions. Here we assume the effect being the same everywhere.

The present study also provides many fruitful avenues for future research. In particular, the following four issues are highly relevant from the Finnish perspective. First, the panel aspect of the available data could be further utilized. For example, Eskelinen and Kuosmanen (2013) apply StoNED method in studying the inter-temporal variation of performance of bank branches. Similar type of an analysis could be informative in the present context to reveal patterns of interruption costs over time and how these patterns are related to the investments on underground cabling. Such information could be helpful for characterizing long term performance targets. Second, it would be interesting to study the whole incentive mechanism itself in more detail. Here we studied only how the reference level for quality should be set. As noted in our description of the Finnish regulatory system, setting the reference level is only one part of the regulation. How the reference level is actually implemented as an incentive mechanism in the regulation is another question, which clearly warrants further investigation. Thirdly, the cost efficiency incentive and the quality incentive are currently modeled as two independent components of the Finnish regulatory model. Clearly, cost efficiency and service quality are intimately related, and it might be preferable to model them jointly. A simple approach is to regulate the total cost that includes the operational costs, capital costs, and interruption costs. The main challenge in this approach is to accommodate the costs of fixed inputs and other nondiscretionary factors that cannot be adjusted by the firm management in the short run. It may be preferable to model the costs of fixed inputs and variable inputs as two separate input factors, imposing the performance targets to the variable inputs only. This requires a model of joint production with multiple inputs and multiple outputs. Dealing with stochastic noise in nonparametric models of joint production remains a methodological challenge. Fourth, as consequence of the previous, the comparison of different ways to measure quality and their consequences on the regulation would warrant a study of its own. Indeed, it would be important to study if the results of regulation change remarkably whether quality and other costs are examined separately as here in this study, jointly within the same model but as separate inputs/outputs, or as an aggregate measure such as SOTEX.

Acknowledgments

The authors gratefully acknowledge funding from the project “Sustainable Transition of European Energy Market (STEEM)” which is part of Aalto University Energy Efficiency (AEF) program. The earlier versions of the study has been presented at the 51st Meeting of the Euro Working Group on Commodities and Financial Modeling (EWGCFM)/ 1st Conference of the Research Centre for Energy Management (RCEM), in London, Great Britain, May 2013, and at the XIII European Workshop on Productivity and Efficiency Analysis (EWEPA) in Helsinki, Finland, June 2013. We are grateful to Heike Wetzel and other participants of these conferences for helpful comments and constructive critique. We also wish to thank the two anonymous reviewers of this journal for their comments.

Please cite this article as: Saastamoinen, A., Kuosmanen, T., Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland, Energy Econ. (2014), http://dx.doi.org/10.1016/j.eneco.2014.04.016

Appendix A. The calculation of the observable interruption costs (EMV, Finnish Energy Market Authority, 2011a)

All prices are in prices of 2005 and they are based on the survey by Silvast et al. (2005). Note that the Finnish EMA uses only the interruptions in middle voltage (1–70 kW) network as the basis for their calculations. Note that the formula accounts both, the unexpected and planned interruptions. Moreover the formula includes both, the duration and number of outages. For duration, the energy based prices are used (subscript *E*), whereas the pricing of number of outages is power based (subscript *W*). Below the term *KA* is similar to the well-known *SAIDI* (System Average Interruption Duration Index) measure and the term *KM* to the *SAIFI* (System Average Interruption Frequency Index) measure. Notice that the report by Silvast et al. (2005) determined the weight of outage time to be significantly larger than the weight for the number of outages. Thus the interruption costs reflect more the costs of long outages than costs of multiple (short) outages. This is desirable as companies are not obligated to compensate very short term interruptions and most of the costs are due to long outages.

$$IC_{t,k} = \begin{pmatrix} KA_{\text{unexp,t}} \times h_{E,\text{unexp}} + KM_{\text{unexp,t}} \times h_{W,\text{unexp}} + \\ KA_{plann,t} \times h_{E,plann} + KM_{plann,t} \times h_{W,plann} + \\ AJK_t \times h_{AJK} + PJK_t \times h_{PJK} \end{pmatrix} \times \left(\frac{W_t}{T_t}\right) \times \left(\frac{CPI_{k-1}}{CPI_{2004}}\right)$$

$IC_{t,k}$	Actual imputed disadvantage caused by electricity supply outages to the DSO's customers in year t in the value of money in year k, euros
$KA_{\text{unexp,t}}$	Customer's average annual outage time weighted by annual energies, caused by unexpected outages in the 1–70 kV network in the year t, hours
$h_{E,\text{unexp}}$	Price of disadvantage caused by unexpected outages to the customer in the 2005 value of money, euros/kilowatt-hour
$KM_{\text{unexp,t}}$	Customer's average annual number of outages weighted by annual energies, caused by unexpected outages in the 1–70 kV network in year t, numbers
$h_{W,\text{unexp}}$	Price of disadvantage caused by unexpected outages to the customer in the 2005 value of money, euros/kilowatt
$KA_{plann,t}$	Customer's average annual outage time weighted by annual energies, caused by planned outages in the 1–70 kV network in year t, hours
$h_{E,plann}$	Price of disadvantage caused by planned outages to the customer in the 2005 value of money, euros/kilowatt-hour
$KM_{plann,t}$	Customer's average annual number of outages weighted by annual energies, caused by planned outages in the 1–70 kV network in year t, numbers
$h_{W,plann}$	Price of disadvantage caused by planned outages to the customer in the 2005 value of money, euros/kilowatt
AJK_t	Customer's average annual outage number weighted by annual energies, caused by time-delayed autoreclosers in the 1–70 kV network in year t, numbers
h_{AJK}	Price of disadvantage caused by time-delayed autoreclosers to the customer in the 2005 value of money, euros/kilowatt
PJK_t	Customer's average annual outage number weighted by annual energies, caused by high-speed autoreclosers in the 1–70 kV network in year t, numbers
h_{PJK}	Price of disadvantage caused by high-speed autoreclosers to the customer in the 2005 value of money, euros/kilowatt
W_t	The amount of energy transmitted to customers from the DSO's electricity network at voltage levels 0.4 kV and 1–70 kV in year t, kilowatt-hours
T_t	number of hours in year t
CPI_{k-1}	consumer price index in year k − 1
CPI_{2004}	consumer price index in year 2004

Appendix B. Online supplementary material

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.eneco.2014.04.016.

References

Ai, C., Martinez, S., Sappington, D.E.M., 2004. Incentive regulation and telecommunications service quality. J. Regul. Econ. 26 (3), 263–285.
Ajodhia, V., 2006. Regulating Beyond Price: Integrated Price–Quality Regulation for Electricity Distribution Networks. Doctoral Dissertation Delft University of Technology.
Ajodhia, V., 2010. Integrated cost and quality benchmarking for electricity distribution using DEA. Int. J. Energy Sect. Manag. 4 (3), 417–433.
Ajodhia, V., Hakvoort, R., 2005. Economic regulation of quality in electricity distribution networks. Util. Policy 13, 211–221.
Alexander, B., 1996. How to construct a service quality index in performance-based ratemaking. Electr. Journal 9, 46–53.
Alvarez, A., Amsler, C., Orea, L., Schmidt, P., 2006. Interpreting and testing the scaling property in models where inefficiency depends on firm characteristics. J. Prod. Anal. 25, 201–212.
Beenstock, M., Goldin, E., Haitovsky, Y., 1998. Response bias in a conjoint analysis of power outages. Energy Econ. 20, 135–156.
Bogetoft, P., Otto, L., 2011. Benchmarking with DEA, SFA, and R. International Series in Operations Research & Management Science, 157. Springer.
Cambini, C., Rondi, L., 2010. Incentive regulation and investment: evidence from European energy utilities. J. Regul. Econ. 38 (1), 1–26.
de Nooij, M., Koopmans, C., Bijvoet, C., 2007. The value of supply security: the costs of power interruptions: economic input for damage reduction and investment in networks. Energy Econ. 29, 277–295.
EMV (Finnish Energy Market Authority), 2011a. Sähkönjakeluverkkotoiminnan ja suurjännitteisen jakeluverkkotoiminnan hinnoittelun kohtuullisuuden valvontamenetelmien suuntaviivat vuosille 2012–2015 (in Finnish, with an English summary as appendix). Available at http://www.emvi.fi (Accessed: 10.4.2013).
EMV (Finnish Energy Market Authority), 2011b. Annual report to the European Commission. Available at: http://www.emvi.fi (Accessed: 4.4.2013).
Eskelinen, J., Kuosmanen, T., 2013. Intertemporal efficiency analysis of sales teams of a bank: stochastic semi-nonparametric approach. J. Bank. Financ. 37 (12), 5163–5175.
Façanha, L.O., Resende, M., 2004. Price cap regulation, incentives and quality: the case of Brazilian telecommunications. Int. J. Prod. Econ. 92, 133–144.
Fan, Y., Qi, L., Weersink, A., 1996. Semiparametric estimation of stochastic production frontier models. J. Bus. Econ. Stat. 14 (4), 460–468.
Fenrick, S.A., Getachew, L., 2012. Cost and reliability comparisons of underground and overhead power lines. Util. Policy 20, 31–37.
Fernandes, C., Candela, A., Gómez, T., 2012. An approach to calibrate incentives for continuity of supply in the Spanish electricity distribution system. Electr. Power Syst. Res. 82, 81–87.
Ferrier, G.D., Trivitt, J.S., 2012. Incorporating quality into the measurement of hospital efficiency: a double DEA approach. J. Prod. Anal. http://dx.doi.org/10.1007/s11123-012-0305-z.
Florens, J.-P., Simar, L., 2005. Parametric approximations of nonparametric frontiers. J. Econ. 124, 91–116.
Fried, H.O., Schmidt, S.S., Yaisawarng, S., 1999. Incorporating the operating environment into a nonparametric measure of technical efficiency. J. Prod. Anal. 12, 249–267.
Fried, H.O., Lovell, C.A.K., Schmidt, S.S., Yaisawarng, S., 2002. Accounting for environmental effects and statistical noise in data envelopment analysis. J. Prod. Anal. 17, 157–174.
Fumagalli, E., Lo Schiavo, L., 2009. Regulating and improving the quality of electricity supply: the case of Italy. Eur. Rev. Energy Mark. 3 (3), 1–27.
Giannakis, D., Jamasb, T., Pollit, M., 2005. Benchmarking and incentive regulation of quality of service: an application to the UK electricity distribution networks. Energy Policy 33, 2256–2271.
Greene, W.H., 2008. Econometric Analysis. Pearson Prentice Hall, New Jersey, USA.
Growitsch, C., Jamasb, T., Pollitt, M., 2009. Quality of service, efficiency andscale in network industries: an analysis of European electricity distribution. Appl. Econ. 41, 2555–2570.
Growitsch, C., Jamasb, T., Müller, C., Wissner, M., 2010. Social cost-efficient service quality — integrating customer valuation in incentive regulation: evidence from the case of Norway. Energy Policy 38, 2536–2544.
Growitsch, C., Jamasb, T., Wetzel, H., 2012. Efficiency effects of observed and unobserved heterogeneity: evidence from Norwegian electricity distribution networks. Energy Econ. 34, 542–548.
Hadri, K., 1999. Estimation of a doubly heteroscedastic stochastic frontier cost function. J. Bus. Econ. Stat. 17 (3), 359–363.
Hadri, K., Guernat, C., Whittaker, J., 2003. Estimation of technical inefficiency effects using panel data and doubly heteroscedastic stochastic production frontiers. Empir. Econ. 28, 203–222.
Hafner, R., Helmer, D., van Til, H., 2010. Investment and regulation: the Dutch experience. Electr. J. 23 (5), 34–46.
Hall, K., 2013. Out of sight, out of mind 2012. An updated study on the undergrounding of overhead power lines. A report prepared for Edison Electric Institute. Available online at: http://www.eei.org/ (Accessed: 17.1.2014).

Please cite this article as: Saastamoinen, A., Kuosmanen, T., Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland, Energy Econ. (2014), http://dx.doi.org/10.1016/j.eneco.2014.04.016

Haney, A.B., Pollitt, M., 2009. Efficiency analysis of energy networks: an international survey of regulators. Energy Policy 37, 5814–5830.
Haney, A.B., Pollitt, M., 2011. Exploring the determinants of "best practice" benchmarking in electricity network regulation. Energy Policy 39, 7739–7746.
Holt, L., 2005. Utility service quality — telecommunications, electricity, water. Util. Policy 13, 189–200.
Jamasb, T., Pollit, M., 2001. Benchmarking and regulation: international electricity experience. Util. Policy 9, 107–130.
Jamasb, T., Pollit, M., 2007. Incentive regulation of electricity distribution networks: lessons of experience from Britain. Energy Policy 35, 6163–6187.
Jamasb, T., Pollit, M., 2008. Security of supply and regulation of energy networks. Energy Policy 36, 4584–4589.
Jamasb, T., Orea, L., Pollitt, M., 2012. Estimating the marginal cost of quality improvements: the case of the UK electricity distribution companies. Energy Econ. 34, 1498–1506.
Johnson, A., Kuosmanen, T., 2011. One-stage estimation of the effects of operational conditions and practices on productive performance: asymptotically normal and efficient, root-n consistent StoNEZD method. J. Prod. Anal. 36 (2), 219–230.
Johnson, A.L., Kuosmanen, T., 2012. One-stage and two-stage DEA estimation of the effects of contextual variables. Eur. J. Oper. Res. 220, 559–570.
Joskow, P.L., 2008. Incentive regulation and its application to electricity networks. Rev. Netw. Econ. 7 (4), 547–560.
Kinnunen, K., 2006. Investment incentives: regulation of the Finnish electricity distribution. Energy Policy 24, 853–862.
Kopsakangas-Savolainen, M., Svento, R., 2008. Estimation of cost-effectiveness of the Finnish electricity distribution utilities. Energy Econ. 30, 209–212.
Kopsakangas-Savolainen, M., Svento, R., 2010. Comparing welfare effects of different regulation schemes: an application to the electricity distribution industry. Energy Policy 38, 7370–7377.
Kopsakangas-Savolainen, M., Svento, R., 2011. Observed and unobserved heterogeneity in stochastic frontier models: an application to the electricity distribution industry. Energy Econ. 33, 304–310.
Korhonen, P., Syrjänen, M., 2003. Evaluation of cost efficiency in Finnish electricity distribution. Ann. Oper. Res. 121, 105–122.
Kuosmanen, T., 2012. Stochastic semi-nonparametric frontier estimation of electricity distribution networks: application of the StoNED method in the Finnish regulatory model. Energy Econ. 34, 2189–2199.
Kuosmanen, T., Kortelainen, M., 2012. Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. J. Prod. Anal. 38 (1), 11–28.
Kuosmanen, T., Saastamoinen, A., Sipiläinen, T., 2013. What is the best practice for benchmark regulation of electricity distribution? Comparison of DEA, SFA and StoNED methods. Energy Policy 61, 740–750.
Kuosmanen, T., Johnson, A., Saastamoinen, A., 2014. Stochastic nonparametric approach to efficiency analysis: a unified framework. In: Zhu, J. (Ed.), Handbook on Data Envelopment Analysis, vol. II. Springer (forthcoming).
Langset, T., Trengereid, F., Samdal, K., Heggset, J., 2001. Quality dependent revenue caps — a model for quality of supply regulation. Proceedings of CIRED 2001. Amsterdam.
Linares, P., Rey, L., 2013. The costs of electricity interruptions in Spain. Are we sending the right signals? Energy Policy 61, 751–760.
McNair, B.J., Bennett, J., Hensner, D.A., Rose, J.M., 2011. Households' willingness to pay for overhead-to-underground conversion of electricity distribution networks. Energy Policy 39, 2560–2567.
Pollitt, M.G., 2005. The role of efficiency estimates in regulatory price reviews: Ofgem's approach to benchmarking electricity networks. Util. Policy 13, 279–288.
Reichl, J., Kollmann, A., Tichler, R., Schneider, F., 2008. The importance of incorporating reliability of supply criteria in a regulatory system of electricity distribution: an empirical analysis for Austria. Energy Policy 36, 3862–3871.
Reichl, J., Schmidthaler, M., Schneider, F., 2013. The value of supply security: the costs of power outages to Austrian households, firms and the public sector. Energy Econ. 36, 256–261.
Sappington, D.E.M., 2005. Regulating service quality: a survey. J. Regul. Econ. 27 (2), 123–154.
Shleifer, A., 1985. A theory of yardstick competition. RAND J. Econ. 16 (3), 319–327.
Silvast, A., Heine, P., Lehtonen, M., Kivikko, K., Mäkinen, A., Järventausta, P., 2005. Sähkönjakelun keskeytyksistä aiheutuva haitta. Helsinki University of Technology and Tampere University of Technology (in Finnish). Available at: http://www.emvi.fi (Accessed: 4.2.2013).
Simab, M., Haghifam, M.R., 2012. Quality performance based regulation through designing reward and penalty scheme for electric distribution companies. Electr. Power Energy Syst. 43, 539–545.
Sullivan, M.J., Suddeth, B.N., Vardell, T., Vojdani, A., 1996. Interruption costs, customer satisfaction and expectations for service reliability. IEEE Trans. Power Syst. 11 (2), 989–995.
Tahvanainen, K., Honkapuro, S., Partanen, J., Viljainen, S., 2012. Experiences of modern rate of return regulation in Finland. Util. Policy 21, 32–39.
Ter-martirosyan, A., Kwoka, J., 2010. Incentive regulation, service quality, and standards in U.S. electricity distribution. J. Regul. Econ. 38 (3), 258–273.
Thakur, T., Deshmukh, S.G., Kaushik, S.C., 2006. Efficiency evaluation of the state owned electricity utilities in India. Energy Policy 34 (17), 1187–1198.
Wang, H.-J., 2002. Heteroscedasticity and non-monotonic efficiency effects of a stochastic frontier model. J. Prod. Anal. 18, 241–253.
Wang, H., Schmidt, P., 2002. One step and two step estimation of the effects of exogenous variables on technical efficiency levels. J. Prod. Anal. 18, 129–144.
White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48 (4), 817–838.
Yu, W., Jamasb, T., Pollitt, M., 2009a. Does weather explain cost and quality performance? An analysis of UK electricity distribution companies. Energy Policy 37, 4177–4188.
Yu, W., Jamasb, T., Pollitt, M., 2009b. Willingness-to-pay for quality service: an application to efficiency analysis of the UK electricity distribution utilities. Energy J. 30 (4), 1–48.

Please cite this article as: Saastamoinen, A., Kuosmanen, T., Quality frontier of electricity distribution: Supply security, best practices, and underground cabling in Finland, Energy Econ. (2014), http://dx.doi.org/10.1016/j.eneco.2014.04.016

ONLINE APPENDIX

Quality Frontier of Electricity Distribution

Antti Saastamoinen, Timo Kuosmanen

Aalto University School of Business
Runeberginkatu 22-24, 00100 Helsinki, Finland
firstname.lastname@aalto.fi
Telephone: +358 9 43131; Fax: +358 9 43138535

Appendix A: *Yearly summary statistics for observed interruption costs (in million €)*

Year	Mean	Std. Dev.	Min	Max	Total
2005	1.467	3.731	0.016	29.4	126
2006	1.465	4.281	0.001	28.8	126
2007	1.290	4.085	0.004	32.3	111
2008	1.306	3.814	0.001	27.5	112
2009	0.982	2.781	0.018	17.9	84.5
2010	2.235	6.483	0.005	44.9	192

n=86 for each year

Appendix B: *Formation of the underground cabling groups*

The underground cabling groups are based on the firm-specific average underground cabling proportion over 2005-2010. The groups are then formed according to the percentile points of the average cabling proportion. The summary statistics are given in Table B below.

Table B: Summary statistic of average underground cabling proportion by firms.

n=86	**Mean**	**Std. dev.**	**Skewness**	**1st quartile**	**Median**	**3rd quartile**	**Min**	**Max**
Mean U.cabling 2005-2010	22.56	27.60	1.25	2.77	8.07	38.18	0.17	100.00

The groups are formed such that the firm belongs in one following groups: <= 1st quartile (low cabling proportion), between 1st quartile and median (medium-low), between median and 3rd quartile (medium), above 3rd quartile (high). The limits of the percentile points are roughly the same as presented in Table 1 since generally the cabling proportion does not change much over the years within one company (i.e. the figures in Table 1 are over all observations, not summary statistics over firm-specific averages as here).

Appendix C: *Estimation results excluding 2 industrial networks*

The sample size used in the estimations of this is 504 as two firms were removed from the sample. Summary statistics of reference values when two outliers are removed are shown in the Table C1 below. Again we pick the value from 2010 as the reference value when using quality frontier. Thus number of observations is 84 in the table below.

Table C1: Summary statistic of the reference values (n=84), millions of €

Variable	Mean	Std. Dev.	Min	Max
EMV IC ref	1.50	4.08	0.02	25.80
Est. IC ref, (1)	1.14	2.16	0.04	13.40
Est. IC ref, (2)	2.17	4.32	0.07	27.90

On average the reference values have increased. This is not surprising as the removed networks had large underground cabling proportions and consequently rather low interruption costs. Compared to the earlier results, on average the reference levels from Model 2 have changed the most. It is likely that one of the removed firms was a frontier firm in the original data. After its removal, a firm with a similar location and cabling level but completely different output profile (and bigger interruptions) is attached with a frontier status. Recall that in Model 1 location in terms of coordinates is excluded and thus frontier is not adjusted with location, only with the level of underground cabling. Although the average level of reference values is somewhat sensitive to the chosen sample/variables, the main result about the stability of reference values between firms is kept intact. This is illustrated in Figure C where the reference values from the quality frontier and EMV approach have been plotted. We see again that the reference values produced by the quality frontier are more stable for companies of similar size than the values from EMA average.

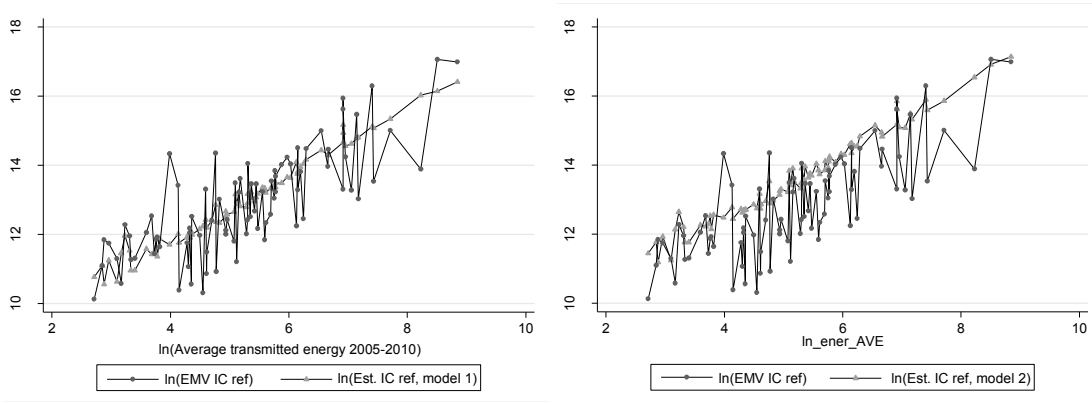

Figure C: The EMV reference levels against the quality frontier reference levels

Lastly we examine how the effects of underground cabling change with the removal of the two firms. Both sets of z-variables have been used (Model 1 & Model 2). The conclusions regarding the level effects of underground cabling remains the same than previously. The level of interruptions is significantly reduced by the underground cabling. Only notable change is in coordinate variables and their significance. The year dummies show the same previously observed pattern.

Table C2: Level effects on the level of interruption costs

	StoNED		2nd stage inefficiency	
	Model 1	Model 2	Model 1	Model 2
U. cabling	-0.028***	-0.025***	-0.023***	-0.024***
	(0.001)	(0.001)	(0.003)	(0.002)
Latitude		-0.038***		0.047**
		(0.014)		(0.023)
Longitude		0.069***		-0.004
		(0.014)		(0.024)
2006	-0.184**	-0.182**	-0.176	-0.185
	(0.090)	(0.087)	(0.130)	(0.119)

2007	-0.329***	-0.330***	-0.332***	-0.350***
	(0.101)	(0.098)	(0.114)	(0.111)
2008	-0.257***	-0.262***	-0.249**	-0.281**
	(0.098)	(0.097)	(0.113)	(0.109)
2009	-0.584***	-0.589***	-0.591***	-0.611***
	(0.104)	(0.101)	(0.109)	(0.106)
2010	-0.126	-0.130	-0.182	-0.197
	(0.125)	(0.119)	(0.104)	(0.101)
Intercept	0.626***	0.595	0.070	-3.260**
	(0.072)	(0.889)	(0.123)	(1.476)

Standard errors in parenthesis, heteroscedasticity robust standard errors for StoNED

*** 1% significance

** 5% significance

Lastly we check the robustness of the variance effects. Analogous to Tables 4 and 5 (in the paper), Tables C3 and C4 present the variance effects in the case of overall error and the decomposed effects. Considering Figure 2 it is not surprising that the significance of the variance effects largely disappears when the 100% cabling company is removed from the sample. It is however notable that the effect is practically zero, not negative and significant as we would expect. This implies that underground cabling does not significantly decrease the variability of interruption costs. The effects on overall error variance do not however yet reveal the effects on the variances of noise and inefficiency. In Table C4 the underground cabling effects on both variance components are insignificant. The coordinate variables are still significant in explaining the inefficiency differentials indicating rather robust location depended performance variation between DSOs.

Table C3: Variance effects on the overall error

	variance effects; StoNED	
	Model 1	Model 2
U. cabling	-0.001	0.000
	(0.002)	(0.002)
Latitude		0.008
		(0.018)
Longitude		0.027)
		(0.019
Intercept	0.502***	-0.749
	(0.057)	(1.179)

Standard errors in parenthesis

*** 1% significance

** 5% significance

Table C4: Decomposed variance effects

	Model 1		Model 2	
	$\boldsymbol{\theta}_u$ inefficiency effects	$\boldsymbol{\gamma}_v$ noise effects	$\boldsymbol{\theta}_u$ inefficiency effects	$\boldsymbol{\gamma}_v$ Noise effects
U. cabling	-0.029	0.003	-0.008	0.004
	(0.026)	(0.003)	(0.010)	(0.004)
Latitude			-0.682***	0.111**
			(0.258)	(0.056)
Longitude			0.443***	-0.044
			(0.156)	(0.068)
Intercept	-0.610**	-1.128***	29.858**	-7.115**
	(0.266)	(0.170)	(14.032)	(3.000)

Standard errors in parenthesis

*** 1% significance

** 5% significance

Productivity analysis is interested in comparing how much more productive for example a firm is than another. Comparison of units can, however, be unfair if the firms operate in highly different environments. For example, a firm might operate in a highly risky environment and thus may look less productive, not necessarily because of any inefficiency, but because of the environment. This dissertation studies what implications a risky operating environment of firms has for productivity and efficiency analysis. The connections between risk and inefficiency are explored with a conceptual discussion and empirical applications. The explicit discussion of these connections is the novel feature of the dissertation. The empirical applications illustrate how risk and operating environment can be interpreted and accounted for in the analysis of aggregate productivity and corruption and the cost-efficiency and quality of service assessments of the Finnish electricity distribution companies.

Aalto-DD 119/2014

ISBN 978-952-60-5808-5
ISBN 978-952-60-5809-2 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Business
Department of Information and Service Economy
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS