Institut d'Administration des Entreprises (IAE) de Lille

École Doctorale SESAM (ED73) : Sciences économiques, Sociales, de l'Aménagement et du Management

Frontier Estimation of Efficiency and Productivity: Some New Perspectives for Firms and Industry

Thèse pour l'obtention du doctorat en sciences de gestion

présentée et soutenue publiquement par

Bouye Ahmed Ould Moulaye Hachem

6 septembre 2011

JURY

Directeur de thèse

Kristiaan Kerstens, Directeur de Recherche au CNRS

Suffragants

Olivier Brandouy, Professeur à l'IAE de l'Université Paris 1

Walter Briec, Professeur à l'Université de Perpignan

Diego Prior, Professeur à l'Universitat Autònoma de Barcelona

Dhafer Saïdane, Maître de conférences (HDR) à l'Université Charles-de-Gaulle

Ignace Van de Woestyne, Professeur à Hogeschool-Universiteit Brussel

L'Université de Lille 1 n'entend donner aucune approbation ou improbation aux opinions émises dans la thèse. Ces opinions doivent être considérées comme propres à l'auteur.

To my mother and to my wife, the best role models one could ask for.

In dedication to my father who passed away much too early.

Estimation de l'efficience et de la productivité : Nouvelles perspectives pour les entreprises et l'industrie

Résumé de la thèse :

Cette thèse se propose d'analyser à la fois théoriquement et empiriquement l'impacte de certaines hypothèses dans la construction des frontières d'efficience sur la productivité et la performance des organisations d'une manière générale, ainsi que d'étudier les mécanismes susceptibles de contrôler leur efficience. Pour atteindre cet objectif, on s'intéresse particulièrement dans nos applications empiriques aux secteurs : bancaire, énergétique, et l'agriculture.

La vocation des frontières d'efficience et de la mesure de productivité est d'évaluer l'inefficience existante dans les unités de production que les dirigeants de l'organisation veulent mettre dans son fonctionnement optimal. Ces méthodes d'évaluation de la performance s'appliquent aussi bien au secteur privé qu'au secteur public et leur utilisation a pour but de comparer le comportement des organisations à travers l'espace et au fil du temps, ou les deux à la fois. C'est ainsi que ces mesures sont utilisées pour analyser l'efficience relative des entités de production (personnes, projets, processus, entreprises, industries, etc.) au sein des grandes organisations (par exemple, des agences bancaires dans une banque). En outre, ces méthodes principales d'analyses comparatives (benchmark) peuvent être utilisées dans une industrie donnée ou à travers des industries, qui peuvent être limitées au nivaux nationale ou international, etc.

Pour ce faire, on constate que les notions d'efficience et de productivité ont été intégrées à différents nivaux dans la formulation de la politique économique et dans différentes

pratiques de gestion. Par exemple, ces mesures d'efficience peuvent être utilisées dans l'application de la réglementation par plafonnement des prix dans les services publics où les organismes de réglementations et les entreprises font grand usage de ces nouvelles techniques pour argumenter le niveau des gains d'efficience, qui doivent être imposées sur les firmes individuelles ou au niveau de l'industrie. De ce fait, l'utilisation de ces méthodes a été intégrée dans les manuels de règlement élaborés par les institutions internationales à l'instar de la Banque mondiale. Les principaux travaux de recherches qui viennent à l'esprit lorsqu'on évoque l'application de ces modèles pour mesurer la performance de différentes organisations sont ceux, entre autres, de Coelli et al. (2003), Jamasb et Pollitt (2001), et Jamasb et al. (2003).

Pour mesurer la performance de l'entreprise par le concept d'efficience, les études de recherche montrent plusieurs méthodes qu'ont été développées et utilisées pour réaliser cet objectif. Littéralement, le concept d'efficience de la production contient différentes approches qui estiment les frontières de production afin d'analyser la performance de l'entreprise par les séries temporelles, à travers les sections, ou par un panel de donnés. Ces approches utilisent les spécifications de frontière de la technologie qui sont estimées par des méthodes paramétriques ou non paramétriques (voir Bogetoft et Otto (2011) ou Coelli et al. (2005)). Dans ce travail de recherche, on adopte l'approche non paramétrique qui permet de déterminer les bonnes pratiques de la technologie par l'enveloppement directe des observations qui sont fixées sur la limite de l'ensemble de production possible. L'estimation non paramétrique de la frontière utilise la programmation mathématique pour envelopper les données aussi étroitement possible sous réserve de certaines hypothèses maintenues. En bref, on se focalise ici sur les deux célèbres méthodes d'analyse qui se basent sur l'approche non paramétrique de frontière. La méthode connue sous le nom FDH («l'enveloppe de libre disposition») et la méthode alternative qui est célèbre sous le nom DEA («Analyse d'Enveloppement de Données»). Dans le cas de FDH la technologie de production suppose seulement une forte disponibilité d'entrées (inputs) et de sorties (outputs), tandis que la méthode DEA ajoute à cette supposition l'hypothèse de convexité et permet ainsi de faire des combinaisons linéaires des unités de productions observés. Briec et al. (2004) ont montré la différence entre les estimateurs de technologie convexe (DEA) et non convexe (FDH), en soulignant que l'impact de la convexité affecte les valeurs des fonctions de coûts et que le deux technologies convexe et non convexe peuvent être combinés avec des différentes suppositions de rendement d'échelle.

Le principal objectif de cette thèse est de parvenir à illustrer l'importance de la convexité dans une grande variété de frontières de production. Afin d'illustrer cet objectif, trois sujets représentent le noyau de cette dissertation : (1) L'utilisation de capacité optimale et la redistribution dans un réseau de branches bancaires allemand ; (2) Une comparaison des indices de productivité Malmquist et Hicks-Moorsteen se concentrant sur l'infaisabilité ; (3) Les économies d'échelle et les rendements d'échelle dans les modèles non paramétriques.

1. La capacité d'utilisation optimale et la réallocation de ressources bancaires: le cas de réseau d'agences bancaires d'une banque Allemande.

Dans ce chapitre, on propose de mesurer la performance opérationnelle de réseaux d'agences bancaires par des nouveaux indicateurs de performance. Pour ce faire, on utilise la notion de capacité d'utilisation sur laquelle se base le modèle industriel de réallocations factorielles, qui permet de déterminer la structure optimale de réseaux d'agences bancaires.

En adoptant l'approche non paramétrique, on cherche à atteindre deux objectifs dans ce chapitre:

1. Montrer comment le modèle de court terme de Johansen peut être utilisé pour développer des scénarios qui permettent la réallocation des entrées (inputs) et sorties (outputs) sur un réseau bancaire afin d'améliorer sa performance.

2. Illustrer comment la convexité peut affecter les résultats du modèle de Johansen à court terme.

En littérature, ce modèle industriel a été utilisé pour mesurer la capacité d'utilisation des entreprises industrielles, car ce modèle propose une possibilité de substitutions entre les inputs et les outputs. C'est ainsi que ce modèle peut intervenir pour optimiser la performance opérationnelle de banques, surtout dans les cas où on ne peut pas changer les inputs fixes à court terme. En outre, ce modèle permet de tracer la frontière des substitutions possibles et le changement de technologie qui en résulte au cours du temps.

L'approche adoptée dans cette étude pour mesurer la performance opérationnelle se sert du modèle industriel qui permet de faire face aux surcapacités des intrants (inputs), particulièrement celles qui sont fixes. L'application empirique est menée sur un échantillon de 142 agences bancaires d'une société financière allemande durant l'année 1998. Le secteur bancaire est un secteur particulièrement intéressant du point de vue de la répartition des centres de décision. Après avoir discuté ces données, nous essayons d'appliquer notre contribution méthodologique et de proposer de nouvelles mesures de performance opérationnelle pour ces agences bancaires. Cette approche a pour objectif de mettre à la disposition des managers en charge du réseau bancaire de nouveaux indicateurs de performance compréhensibles, interprétables et robustes qui leur permettent de prendre les meilleures décisions possibles.

Dans notre approche méthodologique plusieurs mesures de performance ont été introduites et comparées avec des mesures plus classiques comme la mesure de l'efficience technique. Les résultats sont obtenus après avoir calculés les trois principaux indicateurs de performance telle que l'indicateur de l'efficience technique ordinaire, l'indicateur de la capacité d'utilisation et celui de la capacité d'utilisation définie selon Johansen (1972).

Les résultats empiriques montrent que la technologie de production non convexe est plus commode et plus fiable pour détecter l'inefficience chez les agences bancaires que la technologie traditionnelle basée sur la convexité qui est souvent adoptée dans les méthodes d'estimation (benchmark). Le modèle industriel à court terme étant le principal indicateur d'efficience utilisé dans cette étude propose différents scénarios susceptibles d'aider les décideurs à améliorer la performance. Ces scénarios sont bien détaillés dans ce chapitre, ainsi que les intérêts managériaux qui peuvent en être tirés. Ils consistent essentiellement à une fermeture potentielle de certaines agences de ce réseau bancaire avec la préservation des services offerts. Ceci est confirmé par les résultats obtenus avec ce modèle pour le cas de technologie non convexe à l'aide des autres mesures de performances. Plusieurs scenarios supplémentaires sont ajoutés comme autant de conséquences, tel que le transfert des employées et la fixation d'objectifs alternatifs comme l'agrégation des outputs.

Du point de vue empirique, plusieurs limites peuvent être mentionnées. En premier lieu, on constate que les informations liées à l'environnement géographique méritent d'être intégrées dans le modèle. En outre, l'intensité concurrentielle doit être défini et intégré dans ce model, sachant que pour arriver à pallier toutes les limites, le model utilisé peut devenir inapplicable, vue le grand nombre de contraintes mises en place.

On peut finalement dire que l'application de la mesure de l'efficience, l'analyse par la notion de capacité d'utilisation et le modèle industriel à court terme sont des outils fiables est efficaces pour mesurer l'efficience qui permet d'identifier les bonnes pratiques de performance. L'utilité de ce modèle industriel exige qu'il soit intégré dans le système de décisions stratégiques en tant que outil de planification aux niveaux des opérations d'agences bancaires.

2. Comparaison entre deux principaux indices de productivité: Malmquist et Hicks-Moorsteen

L'objectif de ce chapitre est d'expliciter d'un point de vue empirique le problème d'infaisabilité rencontrée souvent lors de l'utilisation des indices de productivité Malmquist, étant un indice de quantité qui mesure dans l'orientation d'inputs ou d'outputs, et Hicks-Moorsteen, qui mesure simultanément une contraction d'inputs et une expansion d'outputs. Ces deux indices qui mesurent le changement de productivité durant différentes périodes se fondent dans leurs définitions sur la fonction de distance définie par Shephard (1970).

L'indice de productivité de Malmquist a été largement utilisé dans la littérature. Cependant cet indice de quantité a montré qu'il n'est pas toujours un indice de changement de productivité. Cet indice peut aussi souffrir de plusieurs infaisabilités lors de son application, qui sont due à ses fonctions de distances d'inputs et d'outputs qui pourraient être parfois indéfinies. Cette infaisabilité est rarement rapportée par les études empiriques, sauf à quelques exceptions à l'instar de travaux de Mukherjee et al. (2001).

En revanche, l'indice de productivité Hicks-Moorsteen a prouvé son aptitude de mesurer le changement de productivité et ceci est dû au fait qu'il est bien défini par les fonctions de distance en inputs et outputs. Hélas, l'utilisation de cet indice dans les travaux de recherche est limitée, bien qu'il soit utilisé par plusieurs travaux de recherche en tant que un bon indicateur de changement de productivité. Il convient à noter que l'indice de Hicks-Moorsteen est dérivé à partir de l'indice de Malmquist que Diewert (1992) avait défini comme un ratio des indices des quantités des outputs et des inputs qui se basent sur les fonctions de distance, cette idée a été inspirée auparavant par Hicks et Moorsteen et développée plus tard par Bjurek (1996).

En général l'utilisation des indices de Malmquist et Hicks-Moorsteen est relativement facile comparée aux autres indices de productivité Fisher et Törnqvist qui ne sont que des cas spéciaux des indices de Malmquist, comme Caves et al. (1982) l'avaient montré, dans la mesure où les deux premier indices ne sont pas très exigeants en terme d'information, par exemple les prix ne sont pas nécessaires. Néanmoins, les deux derniers indices ont l'avantage en terme de programmation comme ils ne nécessitent pas l'estimation des fonctions de distance, mais ils exigent les prix des inputs et des outputs pour être calculés.

Dans ce travail, la comparaison méthodologique entre les deux principaux indices de productivité, Malmquist et Hicks-Moorsteen qui se fondent dans leurs mesures sur la frontière d'efficience, est faite après avoir estimés les deux indices en adoptant une approche non paramétrique basée sur la programmation linéaire. Cette comparaison a pour objectif d'illustrer comment les infaisabilités de l'indice de productivité de Malmquist sont conditionnées par des hypothèses sur la technologie, en particulier (i) l'analyse de court terme par rapport à celle de long terme, (ii) convexe par rapport au non-convexe (iii) le rendement d'échelle constant par rapport à l'hypothèse de rendements d'échelle flexible.

L'application empirique dans ce chapitre est menée sur deux principales catégories de données. La première est un échantillon pris de l'article d'Ivaldi et al (1996). Il s'agit de la quantité de fruit français produite durant trois ans. La deuxième catégorie correspond à la quantité de riz cultivé dans les champs de petits agriculteurs des Philippines, et ce pendant sept années. Après avoir discuté ces données en détail, les mesures de productivité par Malmquist et Hicks-Moorsteen ont été appliquées pour déterminer le changement de productivité sur ces deux différentes activités. Ces différentes mesures ont été estimées selon les deux approches convexes et non convexes, avec rendements d'échelle constants et variables.

C'est ainsi que les résultats de chaque indice sont différents selon l'approche sous jacente adoptée. Plus précisément, l'indice de Malmquist a montré plusieurs infaisabilités dans le cas de rendement d'échelle variable, tandis que l'indice de Hicks-Moorsteen a donné toutes les estimations demandées avec les différentes technologies sans aucune infaisabilité. Par contre le pourcentage d'infaisabilité est très élevé, lorsque la mesure de changement de productivité est prise dans le temps discret par l'indice de Malmquist, cette réalité peut être constaté plus claire aux niveaux des observations individuelles.

Malgré l'inexistence d'infaisabilités dans les mesures des deux indices avec la technologie de rendement d'échelle constant, l'indice de Malmquist a montré qu'il n'est pas toujours un indicateur de changement de productivité adéquat. En revanche, l'indice de Hicks-Moorsteen est apparu comme l'indice le plus pratique pour estimer le changement de productivité en tenant compte les différentes approches.

Bien que cet indice de « Hicks-Moorsteen » soit moins utilisé, il mérite une attention particulière dans une perspective d'utilisation et de développements futurs. Cette conclusion n'exclue pas l'utilisation de l'indice de Malmquist comme mesure fiable du changement de productivité local. Toutefois, ce dernier pouvant être sujet aux infaisabilités, le manager sera en possibilité de recourir à l'indice de Hicks-Moorsteen qui constitue une alternative pérenne de mesure de productivité.

3. Economie d'échelle et le rendement d'échelle dans les modèles non paramétriques : Exploration de l'impact de la convexité

En théorie de production, différentes méthodes non paramétriques sont utilisées pour mesurer l'efficience technique ainsi que l'efficience d'échelle pour des organisations de tout genre. La méthode basée sur l'enveloppe de libre disposition (FDH), tel qu'elle est introduite par Deprins et al. (1984), est conçue pour détendre l'hypothèse de la convexité sur laquelle se base les modèles convexes traditionnels connus sous le nom DEA (Analyse d'Enveloppement de Données).

Afin l'usage potentiel de l'approche FDH, Kerstens et Vanden Eeckaut (1999) ont introduit des hypothèses de rendement d'échelle spécifiques dans sa formulation de base ; ils ont également proposé une nouvelle méthode pour caractériser les rendements d'échelle dans les technologies non-convexes. Briec et al. (2004) ont également développé des fonctions de coût non-convexes qui sont toujours supérieures ou égales à leurs homologues convexes. Ces auteurs ont de surcroit comparé les différentes décompositions traditionnellement convexes avec leurs équivalents non-convexes.

L'idée principale de ce chapitre est d'explorer la différence entre l'efficience technique et l'efficience d'échelle ainsi que les éventuelles différences entre la caractérisation des économies d'échelle et des rendements d'échelle basées sur l'estimation de la fonction de coût et les technologies convexes et non-convexes. Tout ceci a pour objectif d'illustrer la façon dont la convexité de la technologie et la fonction de coût affecte :

1. la décomposition entre l'efficience technique et l'efficience d'échelle dans le contexte des coûts et celui de la production.
2. la caractérisation des économies d'échelle et des rendements d'échelle pour les observations individuelles.

Dans l'analyse empirique, nous utilisons les données de producteurs français de fruits, ainsi que des centrales Chiliennes d'hydroélectriques. Les résultats empiriques montrent la différence entre l'efficience technique et l'efficience d'échelle basées sur des technologies convexes et non-convexes, ainsi que les fonctions de coûts estimées. En outre, ces résultats illustrent les différences entre la caractérisation des économies d'échelle et des rendements d'échelle pour les observations individuelles, suivant les fonctions des coûts convexes et non convexes ainsi que les technologies. Évidemment, pour les observations inefficientes une telle caractérisation est conditionnée par une orientation choisie de mesure.

Il est certain que tout ceci a des conséquences importantes sur les décisions d'investissement et certaines notions clés de l'économie à l'instar de la notion de d'utilisation de capacité.

Mots-clés: Les réseaux d'agences bancaires, Efficience, Capacité, Réallocation, Indice de Productivité de Malmquist, Indice de Productivité de Hicks-Moorsteen, Infaisabilités, Efficience d'Echelle, Rendements d'Echelle, Convexité.

Frontier Estimation of Efficiency and Productivity: Some New Perspectives for Firms and Industry

Summary of the PhD:

This thesis contributes to the efficiency and production frontier literature by adopting a managerial focus to provide a few new solutions to managers. There is in fact one recurrent theme in this PhD: we illustrate the importance of convexity in a wide variety of production frontier modeling settings. Three topics represent the core of this dissertation: (1) Optimal capacity utilization and reallocation in a German bank branch network; (2) A comparison of Malmquist and Hicks-Moorsteen productivity indices focusing on infeasibilities; (3) Scale economies and returns to scale in non-parametric models. The first chapter shows how the short run Johansen model can be used to develop scenarios to manage the reallocation of inputs and outputs over a bank branch network so as to improve its performance and then, we illustrate how convexity affects these results for the short run Johansen model. The second chapter describes how infeasibilities of the Malmquist productivity index are conditioned by assumptions on technology, in particular (i) short-run versus long-run analysis, (ii) convex versus non-convex, and (iii) constant versus flexible returns to scale assumptions. Finally, the third chapter explores the difference between the technical efficiency and the scale efficiency as well as the eventual differences between the characterization of economies of scale and returns to scale based on convex and non-convex technology and cost function estimations.

Keywords: Bank branch network, Efficiency, Capacity, Reallocation, Malmquist Productivity Index, Hicks-Moorsteen Productivity index, Infeasibilities, Scale efficiency, Returns to scale, Convexity.

LEM (UMR 8179)

Lille Economie & Management

École Doctorale SESAM (ED73)

Sciences économiques, Sociales,

de l'Aménagement et du Management

Acknowledgements

The work on this PhD dissertation has been a long journey that started back in 2008 with my arrival in Lille. Overall, the experience has been gratifying in terms of both the insights gained and the skills developed. This rewarding experience has been made possible by the support and enthusiasm of my collaborators and friends.

My debt is evident towards Prof. Dr. Kristiaan Kerstens who accepted being my PhD supervisor after discussing and reorienting my initial project. As an advisor and close collaborator he offered me his expertise and time whenever needed. He opened up the world of academic research in frontier-based estimation of economic and managerial issues and helped me putting my first few steps into this world.

I wish to express my deep appreciation to all my professors during my years of study and especially the colleagues at LEM (UMR 8179) who offered their professional and moral guidance throughout the doctoral program at the university of Lille 1. Prof. Dr. Alain Desreumaux, director of the doctoral school SESAM (ED73), deserves special thanks for guiding me through the administrative procedures related to obtaining the PhD.

I am grateful for the small monthly grant from the Ministry of Higher Education from Mauritania, my home country over the years. However, the lack of sufficient funding has not made my personal life any easier while working my way through the PhD program. I have been obliged to make up for the difference to cover my cost of living modestly in Lille. In any case, while the PhD did not make me any richer (yet), it did make me much stronger in terms of a person and an intellectual. That is probably what a PhD is all about in the first place.

Furthermore, I am especially grateful to my friends in Lille and at home. I highly appreciate their friendship and continuous encouragement. Finally, I would like to thank my

mother, my wife and my family. Their love and wisdom have been a constant source of inspiration and strength throughout all my years of study. I therefore dedicate this thesis to them.

Table of Contents

RESUME DE LA THESE : 4

SUMMARY OF THE PHD: 13

GENERAL INTRODUCTION 21

1. REVIEW OF THE LITERATURE AND RECENT RESEARCH DEVELOPMENTS 21

2. OBJECTIVES OF THE STUDY, METHODS, AND ACHIEVEMENTS 28

2.1. OPTIMAL CAPACITY UTILIZATION AND REALLOCATION IN A GERMAN BANK BRANCH NETWORK 28
2.2. MALMQUIST AND HICKS-MOORSTEEN PRODUCTIVITY INDICES AND INFEASIBILITIES 30
2.3. SCALE ECONOMIES AND RETURNS TO SCALE IN NON-PARAMETRIC MODELS: EXPLORING THE IMPACT OF CONVEXITY 31

CHAPTER 1: OPTIMAL CAPACITY UTILIZATION AND REALLOCATION IN A GERMAN BANK BRANCH NETWORK: EXPLORING SOME STRATEGIC SCENARIOS 33

1. INTRODUCTION 34

2. METHODOLOGY 40

2.1. INTRODUCTION 40
2.2. DEFINITIONS OF EFFICIENCY, PLANT CAPACITY, AND THE SHORT-RUN INDUSTRY MODEL .. 43
2.3. SHORT-RUN INDUSTRY MODEL: ADDITIONAL SCENARIOS 48

3. DATA: BANK BRANCHES OF A GERMAN SAVINGS BANK 51

4 EMPIRICAL RESULTS 54

4.1. ESTIMATION OF PLANT CAPACITY: TESTING FOR CONVEXITY 54
4.2. SHORT-RUN INDUSTRY MODEL: BASIC RESULTS AND ADDITIONAL SCENARIOS 56
4.3. RESULTS FOR INDIVIDUAL BANK BRANCHES: SOME EXAMPLES 62

5. CONCLUSIONS 64

CHAPTER 2: MALMQUIST AND HICKS-MOORSTEEN PRODUCTIVITY INDICES: AN EMPIRICAL COMPARISON FOCUSING ON INFEASIBILITIES 73

1. INTRODUCTION 74

2. DEFINITIONS OF TECHNOLOGY AND PRODUCTIVITY INDICES 77

2.1 TECHNOLOGY AND DISTANCE FUNCTIONS 77
2.2 MALMQUIST AND HICKS–MOORSTEEN PRODUCTIVITY INDICES 78
2.3 SHORT-RUN MALMQUIST AND HICKS–MOORSTEEN PRODUCTIVITY INDICES 80
2.4. INFEASIBILITIES IN THE LITERATURE: A SELECTION 81

3. METHODOLOGY: SPECIFICATION OF TECHNOLOGIES 84

4. SAMPLE DESCRIPTIONS **86**

5. EMPIRICAL RESULTS **87**

6. CONCLUSIONS **91**

APPENDIX: COMPUTING MALMQUIST AND HICKS-MOORSTEEN INDICES: LINEAR PROGRAMMING PROBLEMS **93**

CHAPTER 3: SCALE ECONOMIES AND RETURNS TO SCALE IN NON-PARAMETRIC MODELS: EXPLORING THE IMPACT OF CONVEXITY **102**

1. INTRODUCTION **103**

2. TECHNOLOGY, COST FUNCTION AND EFFICIENCY DECOMPOSITION **105**

3. TECHNOLOGY AND COST FUNCTION SPECIFICATIONS **111**

4. DESCRIPTION OF THE SAMPLES **115**

5. EMPIRICAL RESULTS **116**

6. CONCLUSIONS **120**

GENERAL CONCLUSIONS **126**

1. KEY CONCLUSIONS BY CHAPTER **126**

1.1 MANAGING A BANK BRANCH NETWORK'S PERFORMANCE 126

1.2. MALMQUIST AND HICKS-MOORSTEEN PRODUCTIVITY INDICES: DOCUMENTING INFEASIBILITIES 128

1.3 SCALE ECONOMIES AND RETURNS TO SCALE IN NON-PARAMETRIC MODELS 128

2. LIMITATIONS AND AVENUES FOR FUTURE RESEARCH **129**

BIBLIOGRAPHY **130**

List of Tables

Chapter 1

Table 1: Descriptive Statistics of Inputs and Outputs

Table 2: Descriptive Statistics for θ_1, θ_2 and CU_{eo}

Table 3: Li (1996) Test Statistic for Differences in Densities

Table 4: Descriptive Statistics of Plant Capacity Inputs and Outputs: Convex vs. Non-Convex; Full Efficiency vs. Full Inefficiency

Table 5: Basic Short-Run Industry Model Results: Impact of Convexity and Technical (In)efficiency

Table 6: Short-Run Industry Model Results: Additional Scenarios

Table 7: Results for Individual Bank Branches: Some Examples

Chapter 2

Table 1: Input-Oriented Malmquist vs. Hicks–Moorsteen Indices: Descriptive Statistics

Table 2: Input-Oriented Malmquist: Descriptive Statistics of Infeasibilities Across Periods

Table 3: Input-Oriented Malmquist: Descriptive Statistics of Infeasibilities Across Units

Table 4: Input-Oriented Malmquist vs. Hicks–Moorsteen Indices: Rank Correlations

Chapter 3

Table 1: Non-Convex and Convex Decompositions

Table 2: Spearman Rank Correlation Coefficients between Convex and Non-Convex Decomposition Components and between Production and Cost Perspectives

Table 3: Returns to Scale and Economies of Scale Results

Table 4: Returns and Economies of Scale: Differences between Convex and Non-Convex Methods

Table 5: Returns and Economies of Scale: Differences between Production and Cost Models

List of Figures

Chapter 1

Figure 1a: Industry Efficiency measure in Relation to α in Convex Case

Figure 1b: Industry Efficiency measure in Relation to α in Non-convex Case

Chapter 2

Figure 1: Input-Oriented Malmquist Index: Infeasibilities and Technology Assumptions

Figure 2: Input-Output Section With Position of Observation 18 (Coelli et al. (2005))

Figure 3: Subvector Malmquist and Hicks-Moorsteen for Observation 35 (Coelli et al. (2005))

General Introduction

1. REVIEW OF THE LITERATURE AND RECENT RESEARCH DEVELOPMENTS

The last two decades have witnessed an unprecedented interest in efficiency measurement in the academic literature. This has also led to a massive amount of publications reporting on the efficiency and productivity growth measures in a wide variety of industries in the private sector as well as in the public sector. Examples of well-studied sectors for which surveys of the literature are available include banking (see Hughes and Mester (2008)), education (Worthington (2001)), health care (Ozcan (2008)), insurance (Cummins and Weiss (2000)), railways (Nash and Smith (2008)), water industry (Abbott and Cohen (2009)), among others.

But, efficiency measurement and derived measures of productivity growth have also entered the domain of policy formulation and practical management at various levels. One can think first and foremost about the role of efficiency measurement in implementing price cap regulation in utilities, where both regulators and firms make ample use of these new techniques to argue about the level of the efficiency gains to be imposed on individual forms or the industry level. The use frontier-based efficiency and productivity notions has become integrated in the handbooks for regulation developed by international institutions like the World Bank (e.g., Coelli et al. (2003)). Jamasb and Pollitt (2001) offer a survey on the countries implementing such performance benchmarking schemes in the electricity sector, while Jamasb et al. (2003) discuss survey results from regulators on the gaming issues that can occur when implementing such yardstick competition.

In addition, efficiency and productivity measurement has meanwhile found its way to management in certain industries. Because of evident reasons of discretion and secrecy, such implementations of frontier benchmarking are much less documented in published sources. Therefore, we just mention two documented managerial applications.

First, Fried et al. (1995) document how the trade association of USA credit unions (CUNA) has got involved in an almost continuous effort to deliver its members monitoring services to control the evolution of their efficiency and productivity using simple to understand non-convex frontier models. While access restrictions temper competition among credit unions, these frontier models help their members improving their strategic positioning relative to commercial banks and other financial institutions, with whom they are in competition at the local level.

Second, Sherman and Ladino (1995) describe how a small regional bank in the USA with a limited network has managed to use traditional frontier benchmarking models to realize substantial cost savings to generate the necessary internal financing for a major expansion strategy. Combined with field visits and an activity analysis by questionnaires, this bank in the end managed to economize $6 million of the potential $9 million total expenses indicated by the frontier methods as being wasted.

Benchmarking analysis represents a key tool in business economics. Two main activities of any manager controlling and supervising an organization are on the one hand monitoring (assessing how the firm is doing over time), and benchmarking (comparing firm performance with respect to its main competitors). While both activities aim to enhance performance, monitoring is clearly oriented towards the internal organization, while benchmarking by definition takes an external perspective (e.g., Balk (2003)).

Benchmarking can ideally be conceived as the search for and the emulation of the best practices in an industry. Through benchmarking, a firm can deduce whether it has managed to adapt best or worst practices. Knowing its relative performance, it can target to maintain its eventual superiority or to close the eventual gap relative to its best practice competitors (Camp (1998)). While benchmarking is probably an age-old idea, it's origin is nowadays often related to the activities of Xerox in 1980 when it compared its photocopier production in the USA with those of Fuji-Xerox in Japan. This has been quickly followed by a widespread adaptation by firms in search for improvements (Voss et al. (1997)).

Traditionally, organizations in different industries have used some simplified efficiency or productivity measures (e.g., partial productivity measures, also sometimes known as Key Performance Indicators (KPIs)) as a basic benchmarking tool to assess the efficiency and effectiveness of firms. While these tools offer a myriad of ratios or indicators indicating some aspect of performance, it is often hard if not impossible to aggregate this variety of ratios or indicators into some summary measure. This way of tackling measurement problems in production is often based on the strong tradition to use a variety of ratios in accounting, business and finance (see, e.g., Bragg (2002) for an overview).

More recently, some managers and policy makers have discovered that inefficiencies can be better identified by using frontier-based efficiency or productivity measures, since these allow taking into account the multidimensional nature of modern production technologies in industries and services. The use of well-defined efficiency measures has the double advantage of offering a simple aggregate indication of performance over several dimensions and having a measure with a meaningful economic interpretation. For instance, input-oriented efficiency measures normally have a cost interpretation which allows results to be immediately translated into budgetary policies.

These methods of performance assessment apply to the private as well as to the public sector. Furthermore, these methods are relevant for both the production of goods and services.

The purpose of frontier-based efficiency or productivity measures is to compare the behavior of organizations over time, across space, or both. Furthermore, this performance measurement can be used to analyze the relative efficiency of production entities (individuals, projects, processes, firms, industries, etc.) within larger organizations (e.g., bank branches within a bank). Furthermore, such benchmarking comparisons can be made within a given industry or across industries, can be limited to the national level or may even have an international character, etc.

We first quickly review the most relevant efficiency and productivity notions and explain how to make them operational. Farrell (1957) introduced the idea of best-practice frontiers and provided the first measurement scheme for efficiency by distinguishing between technical and allocative efficiency. More recent work has created more elaborate taxonomies by adding scale and structural efficiencies (see, e.g., Färe et al. (1994)).

First, technical efficiency requires production to be situated on the boundary of the production possibility set or technology. Technology summarizes all possibilities of transforming inputs into outputs. A producer is technically inefficient if production occurs in the interior of technology. Second, structural efficiency is a special case of technical efficiency. A technically efficient producer is structurally efficient if production is situated in the economic (uncongested) region of production. Structural inefficiency happens when production faces some form of congestion, whereby some of the inputs have negative marginal products. Third, scale efficiency measures the eventual divergence between the actual and ideal scale of production. The ideal scale of production is represented by a long run competitive equilibrium situation, whereby production satisfies constant returns to scale. An organization is scale efficient if its production occurs on a constant returns to scale frontier. Otherwise, it is scale inefficient. Finally, allocative efficiency is defined by a point on the boundary of technology satisfying a given objective of the organization, given certain constraints on prices and quantities. Organizations are often assumed to be minimizing costs. A technically efficient producer is then allocatively inefficient if observed costs are situated above minimal costs.

To estimate production frontiers, a variety of methods have been developed for analyzing time series, cross-section or panel data. For convenience, this discussion focuses on cross-section data. Once frontiers are estimated, productivity changes can be derived from measuring the shifts in the frontier over time. To measure performance, the literature contains several approaches that can be used to specify and measure the efficiency of firms. These approaches employ frontier specifications of technology that are estimated by either parametric or non-parametric methods:

- The parametric approach assumes that the boundary of technology can be represented by a particular functional form with a finite number of parameters;
- The non-parametric approach directly envelops the observations in the sample while imposing minimal regularity axioms on the technology.

These methods are different from one another in terms of the underlying behavioral assumptions and data requirements. But more importantly, these methods differ to the extent that these allow for random error or not:

- Stochastic methods allowing for measurement error in addition to inefficiency;
- Deterministic methods assume that observations are measured without error.

Combining both of these distinctions yields a four-way classification. Surveys of all methods confounded are available in Bogetoft and Otto (2011) and Coelli et al. (2005). The present PhD focuses on non-parametric deterministic approaches and ignores the other methods. Specialized surveys focusing on the non-parametric deterministic frontier estimation methods are Ray (2004) and Thanassoulis et al. (2008), among others.

Deterministic non-parametric methods obtain a best practice technology by directly enveloping the observations on the boundary of the production possibility set. These extremal estimators use mathematical programming to envelop the data as tightly as possible subject to certain maintained assumptions. These production assumptions are far less restrictive than the ones used in other approaches. We briefly present two families of important technologies. A production technology only assuming strong input and output disposability is known as the free disposal hull (FDH). Strong input disposability implies that any given level of outputs remains feasible if any of the inputs increases. Strong output disposability means that it is always possible to reduce outputs with given inputs. An alternative production technology adds convexity to the assumptions maintained by FDH. Convex non-parametric frontiers, known as Data Envelopment Analysis (DEA) models, allow for linear combinations of observed production units. Briec et al. (2004) have shown that the convex technology estimators indicate larger or equal amounts of inefficiency compared the non-convex technology and especially that the impact of convexity also affects the values of cost functions: convex cost functions indicate lower or equal values compares to non-convex cost functions. This observation is related to the properties of the cost function in the outputs. Both convex and non-convex families of technologies can be combined with various assumptions regarding returns to scale (see Briec et al. (2004)).

Many of analysts in the economics and management literature have looked at the notion of capacity utilization and its ability to help assessing the performance of organizations. In the theoretical production literature, the capacity notion comes in at least two varieties: a technical (engineering) and an economic capacity concept (see, e.g., Johansen (1968) and Nelson (1989)). The former concept of capacity utilization considers only the physical information on inputs and outputs, while the later notion includes the price information and founds its measure mostly on the cost function.

A bit surprisingly, the majority of the frontier estimation literature has –often implicitly– taken a long run perspective (assuming that all inputs and/or outputs are under managerial control) and has almost completely ignored the capacity utilisation notion as a bridge concept between a short-run and a long-run analysis.

Johansen (1968) defined capacity utilization by using a non-parametric production frontier indicating the maximum potential output that could be produced per unit of time with existing plant and equipment, assuming no restriction of variable inputs. This capacity notion has been used in several studies applying this concept at the industry level (e.g., Coelli et al. (2002)).

Finally, while the overview so far has been mainly limited to a static evaluation of performance, it is equally important to be able to evaluate performance over time. A variety of productivity indices aim to measure the total factor productivity (TFP) of firms and sectors. Rather famous productivity indexes widely used by economic analysts are the Malmquist and Hicks-Moorsteen productivity indices on the one hand and the Fisher and Törnqvist productivity indices on the other hand. Both the Malmquist and Hicks-Moorsteen productivity indexes do not require input and output prices, but require a detailed knowledge of the technology obtained from combining a variety of efficiency measures. By contrast, the other productivity indexes need data on input and output prices to aggregate the quantity information on inputs and outputs, but they have to make stronger assumptions than the two former productivity indexes (e.g., assume allocative efficiency). These different productivity measures have been surveyed by, for instance, Diewert (1992) or Diewert and Nakamura (2003).

2. OBJECTIVES OF THE STUDY, METHODS, AND ACHIEVEMENTS

There is in fact one recurrent theme in this PhD: to illustrate the importance of convexity in a wide variety of production frontier modeling settings. The next subsections describe in more details how this principal objective is articulated into the three main chapters. Each of these chapters aims to fill up some gaps in the literature as well as to provide a few new solutions to managers.

2.1. Optimal capacity utilization and reallocation in a German Bank Branch Network

Bank management traditionally monitors the operational performance of its branch network by a variety of tools. Massive amounts of studies have assessed banking efficiency. Fewer studies have considered bank branch network efficiency, mostly within a single bank network. Studies analyzing retail banking efficiency using frontier methods have amply shown how bank management can monitor the operational efficiency of its central services and of its branch network by means of these new tools (see, e.g., Oral and Yolalan (1990)). Rather common managerial uses of frontier benchmarking results for the management of a branch network include: efficiency measures can be used to set cost and revenue targets, to identify branches in need of an internal audit, to select the best branches to train new employees to adopt best practices, to induce learning practices by putting together weak and strong performers, etc.

While the measurement of the efficiency of bank branch networks has become fairly standard, few managerial tools are available to optimize existing bank branch networks while somehow correcting for existing inefficiencies and accounting for targets, common resource constraints, and policy concerns of various kinds. One can find a small literature that starts from efficiency measures at the firm level to come up with some reallocation of resources at the industry or network level. Examples among a large variety of research proposals include the articles by Athanassopoulos (1995), Golany and Tamir (1995), Lozano and Villa (2004), among others. To our knowledge, no article so far has managed to discover a common structure in these proposals.

To start from firm foundations, we have therefore in opted to stick close to a variation of the short-run industry model initially proposed in Johansen (1972). This model received quite a bit of discussion in the economics literature at the time (e.g., Førsund and Hjalmarsson (1983)). In particular, Dervaux et al. (2000) have linked this model to the frontier-literature by introducing frontier-based estimates of capacity utilization. This refined short-run industry model has been adapted to analyze excess capacities in fisheries while offering a large choice of policy options in Kerstens et al. (2006).

The chapter makes use of the sample of 142 German bank branches in the year 1998 described in Porembski et al. (2005), the only database on a banking network we could lay our hands on. Given this origin of our data, we have no price information, which limits our choice of capacity utilization concepts drastically. Instead of measuring some cost-based capacity notion, we are forced to settle for a plant capacity notion (see Johansen (1968)).

This chapter's specific objectives are to:

(i) show how the short run Johansen model can be used to develop scenarios to manage the reallocation of inputs and outputs over a bank network so as to improve its performance, and

(ii) illustrate how convexity affects the results from the short run Johansen model.

To the best of our knowledge, these goals have not been illustrated in the existing literature.

This chapter is joint work and has been published in the following book:

Kerstens, K., B.A. Moulaye Hachem, I. Van De Woestyne, N. Vestergaard (2010) Optimal Capacity Utilization and Reallocation in a German Bank Branch Network:

Exploring Some Strategic Scenarios, in: A. Tavidze (ed) *Progress in Economics Research* (Volume 16), New York, Nova Science, p. 35-61.

Notice that subsection 4.3 is new and has not been part of this publication.

2.2. Malmquist and Hicks-Moorsteen Productivity Indices and Infeasibilities

Frontier efficiency methods have been widely used to obtain estimates of total factor productivity indices in a discrete time framework. Especially the Malmquist productivity index has been very popular, among others since it allows to decompose dynamic performance into technical change (movements of the production frontier) on the one hand and changes in technical efficiency (changes in the relative positioning with respect to the moving production frontier) on the other hand. Some studies have reported that this index shows some infeasibilities caused by its undefined underlying output or input distance functions (the inverses of the corresponding radial efficiency measures). By contrast, the less popular Hicks-Moorsteen productivity index is known to be feasible under standard assumptions.

While managers are supposed to care most about profits, they should also take great interest in productivity change. Indeed, it has been argued that changes in total factor productivity (*TFP*) are the main impetus for changes in profits. Balk (2003: 6) has phrased this aptly as follows: “the most encompassing measure of productivity change, *TFP* change, is nothing but the ‘‘real’’ component of profitability change. Put otherwise, if there is no effect of prices then productivity change would coincide with profitability change.” Therefore, the continuous monitoring of the evolution of productivity should be an implicit if not an explicit concern for managers. The absence of such information due to infeasibilities complicates this task.

This chapter has the specific goal to:

illustrate how infeasibilities of the Malmquist productivity index are conditioned by assumptions on technology, in particular (i) short-run versus long-run analysis,

(ii) convex versus non-convex, and

(iii) constant versus flexible returns to scale assumptions.

These goals have not been systematically explored in the existing literature. In the empirical part of the chapter, we apply these theoretical issues on the agriculture sector by taking two panel data sets available from published sources. The first data contains an unbalanced panel of three years of French fruit producers, while the second data set concerns a small balanced panel of smallholder rice farmer in the Tarlac region of the Philippines observed over seven years.

This chapter is joint work and an earlier version has been published as a discussion paper:

Kerstens, K., B.A. Moulaye Hachem, I. Van De Woestyne (2010) Malmquist and Hicks-Moorsteen Productivity Indices: An Empirical Comparison Focusing on Infeasibilities, Lille, LEM (Document de travail 2010-09), 25 pp.

Notice that this chapter is currently under review.

2.3. Scale Economies and Returns to Scale in Non-Parametric Models: Exploring the Impact of Convexity

Non-parametric frontier efficiency methods are popular estimate technical and scale efficiencies of various organisations. The Free Disposal Hull (FDH) imposes a free disposal assumption on inputs and outputs, but relaxes the convexity assumption underlying the traditional convex models. Kerstens and Vanden Eeckaut (1999) introduce specific returns to scale assumptions into the basic FDH model and propose a new goodness-of-fit method to infer the characterization of returns to scale for non-convex technologies. Finally, Briec et al. (2004).

also propose non-convex cost functions that are always larger or equal to their convex counterparts and they relate the traditional convex decomposition into technical, scale, allocative and overall efficiency with its non-convex counterparts.

This chapter has the specific goal to:

illustrate how convexity on technology and cost function affects

(i) the decomposition between technical and scale efficiencies in a production and cost context,

(ii) the characterization of economies of scale and returns to scale for individual observations.

Again, these goals have –to the best of our knowledge- not been systematically explored in the economic or operations research literatures. In this chapter we explore these differences between technical and scale efficiencies based on both traditional convex and these rather new non-convex technology and cost function estimations. In the literature, estimates based on the cost function have never been reported. Equally important, we also point to the eventual differences between the characterization of economies of scale and returns to scale for convex as well as non-convex cost functions and technologies for individual observations. For the purpose of empirical illustration, we apply these efficiency decomposition on the agriculture sector and on the hydro electricity industry. The first data contains an unbalanced panel of three years of French fruit producers, the other hydro-electric power generation plant data are observed on a monthly basis for a single year.

This chapter is joint work with the PhD supervisor.

Chapter 1:

Optimal Capacity Utilization and Reallocation in a German Bank Branch Network: Exploring Some Strategic Scenarios

Abstract:

Quite a few studies have considered efficiency at the bank branch level by comparing mostly a single branch network, while an abundance of studies have focused on comparing banking institutions. However, to the best of our knowledge no study has ever assessed performance at the level of the branch bank network by looking for ways to reallocate resources such that overall performance improves. Here, we introduce the Johansen-Färe measure of plant capacity of the firm into a multi-output, frontier-based version of the short-run Johansen industry model. The first stage capacity model carefully checks for the impact of the convexity assumption on the estimated capacity utilization results. Policy scenarios considered for the short-run Johansen industry model vary in terms of their tolerance with respect to existing bank branch inefficiencies, the formulation of closure policies, the reallocation of labor in terms of integer units, etc. The application to a network of 142 bank branches of a German savings bank in the year 1998 measures their efficiency and capacity utilization and demonstrate that by this industry model approach one can improve the performance of the whole branch network.

Keywords: Bank Branch Network, Efficiency, Capacity, Reallocation

JEL classification: G21, M11.

1. INTRODUCTION

In today's integrated financial markets, banks face increasing competition for market share. The rapid changes in market conditions (e.g., disintermediation and deregulation trends, successive merger waves, new competition from the non-financial sector) raise a number of important questions from a regulatory perspective about the structure of the banking industry. But, equally important are the strategic issues related to the management of these financial service providers offering a wide range of increasingly complex products. Against this background, the issue of bank efficiency has become rather prominent, since inefficient banks may not survive these continuous challenges, especially when the sector implements massive investments in IT to foster productivity growth (improved information management, new delivery channels, etc.). While the literature on the efficiencies of banking institutions has been summarized from various perspectives (see, among others, Berger (2007), Berger and Humphrey (1997), Goddard et al. (2001), and the focused surveys on consolidation of Amel et al. (2004) and Berger et al. (1999)), the literature analyzing the drivers of performance in financial services delivery remains rather limited (see Harker and Zenios (2001)) as does the literature on the management of bank branch networks (see Paradi et al. (2004) for a survey).

An abundant amount of studies has focused on comparing banking institutions, while fewer studies have studied efficiency at the bank branch level by comparing mostly a single branch network. However, to the best of our knowledge no study has ever assessed the performance at the level of the branch bank network by looking for ways to reallocate resources such that overall performance of the network improves. To put this topic in perspective, we first briefly summarize the efficiency literature on banking institutions and bank branch networks. Then, we expand on the reasons why the management of a branch network requires new models and how the short-run Johansen industry model shows some promise in this respect.

In view of the dual role of financial institutions as providers of transactions and as intermediates transferring funds from savers to investors, in the efficiency literature one finds mainly two types of models to measure the flow of services in a given period (see Berger and Humphrey (1997)):

- Production approach: Banks are considered as service providers to account holders that perform transactions and process documents for depositors (e.g., checks, loan applications, credit reports, etc.). Outputs are defined in terms of numbers of transactions or documents processed. Only current expenses related to physical inputs like labor and capital and their associated costs are considered, while interest payments are ignored. As a consequence, only input prices for physical inputs are considered.
- Intermediation approach: Banks are intermediating funds between savers and investors. The flow of services is seen as proportional to the stock of financial value in the accounts (e.g., value of loans, deposits, etc.). Outputs are defined in terms of financial value terms. In addition to the physical inputs, also the input of funds is considered. Costs therefore contain current expenses and interest payments and input prices for physical inputs and financial inputs are taken into account.

Both approaches have their relative advantages (see again Berger and Humphrey (1997)). The intermediation approach is more appropriate for evaluating entire banking institutions, since interest expenses are an important part of total costs and need to be minimized to guarantee overall cost minimization or profit maximization. The production approach is most suitable for bank branches, since intermediation is organized at a higher level. Certain studies employ both approaches.

Since the seminal article of Berger et al. (1997), some progress had been made in analyzing bank branch efficiency. Some key results from this limited literature can be summarized as follows. (i) There are scale economies at the branch level. But, the excess costs of over-branching are rather low due to the relative flatness of average cost curves. Furthermore, additional revenues gained from the convenience offered to the customers at the network level probably compensate these additional costs due to scale inefficiency. (ii) The large dispersion of technical inefficiencies at the branch level implies that technical inefficiencies at the bank level are understated, since even efficient banks are likely to have some inefficient branches. (iii) Bank management only imperfectly controls the costs of branch offices through its procedures, incentives and supervision. The quality of local management remains a crucial determinant of branch performance. Further conclusions on bank branch efficiency are found in the surveys of Berger and Humphrey (1997) and Paradi et al. (2004). International comparative network studies are still extremely rare (see Athanassopoulos et al. (2001) or McEachern and Paradi (2007) for exceptions).

Bank management has always monitored the operational efficiency of its branch network by a variety of tools to measure its performance. Traditional tools to measure efficiency are based on financial ratios (such as Return on Assets, Return on Equity, or similar ratios). While ratios provide a great deal of information about financial performance in comparisons across time or relative to other banks' performance, these tools have well-known limitations. An alternative approach is the use of deterministic or econometric frontier efficiency analysis using a production approach or eventually using accounting information (as it turns out that financial and production performance tends to be rather correlated: see, e.g., Elyasiani et al. (1994) or Feroz, Kim and Raab (2003)). Some success stories of using frontier benchmarking in evaluating branch networks have been well-documented (see, e.g., Sherman and Ladino (1995) or Athanassopoulos and Giokas (2000)). Straightforward uses of frontier benchmarking for managing branch networks have equally been testified in a variety of written sources. In particular, efficiency scores, rankings and frontier projections have, among others, been used as an instrument to reformulate budgetary and revenue targets; to identify branches needing a thorough internal audit; to rewrite internal procedures and test the implications of these reforms on performance; to induce a learning process for current personnel by assembling both weak and good performers and eventually move best-practice managers to poor performing branches; to train new employees at best practice branches, etc.

However, the rapid technological changes have led to the introduction of new delivery systems (Automatic Teller Machines (ATM), electronic fund transfer of point of sale (EFTPOS), phone and internet banking, e-money, centralized call centers, etc) that risk to erode away the earlier dominance of the brick-and-mortar bank branch. This increasing competition of distribution channels goes hand in hand with an increasing number of bank branches in the USA (Thirtle (2007)), even though these branches are becoming more concentrated in the networks of just a few institutions (due to industry consolidation). Though Thirtle (2007) finds no systematic relationship between branch network size and overall institutional profitability, which seems to suggest that banks somehow optimize the size of their branch network as part of an overall strategy, her findings do suggest that banks with mid-sized branch networks (101–500 branches) may be at a competitive disadvantage in branching activities relative to banks with larger branch networks. Together with the common knowledge that there remain unexploited scale economies at the branch level whereby the additional cost of "overbranching" seems to be compensated by the gains in additional revenues from providing extra customer convenience (see above), these findings point to the conclusion that the management of branch networks is going to remain a major challenge for the years to come.

While measuring the efficiency of bank branch networks is fairly standard, few if any managerial tools are available to optimize existing bank branch networks while correcting for existing inefficiencies and accounting for targets of various kinds. A burgeoning literature exists that starts from efficiency measurements at the individual firm (plant or subunit) level to come up with some reallocation of resources at the level of the industry (firm). Early examples of such articles include Athanassopoulos (1995), Färe et al. (1992), Golany and Tamir (1995), Li and Ng (1995), among others. Meanwhile, a series of additional publications have appeared, including, for instance, Asmild et al. (2009), Giménez-García et al. (2007), Korhonen and Syrjänen (2004), and Lozano and Villa (2004). However, it is difficult to see a common structure in this large variety of research proposals. Furthermore, since few empirical applications exist and experience with practical implementations seems absent (at least it is not reported in publications), it is difficult to assess the relative advantages of these models from a managerial viewpoint. To the best of our knowledge, none of these reallocation models has ever been applied to the banking sector.

We have therefore opted to stick to a short-run industry model initially proposed in Johansen (1972) which received at least a minimum of discussion in the economics literature (see, e.g., Førsund and Hjalmarsson (1983) or Hildenbrand (1981)). Furthermore, it has been linked to the frontier-literature in Dervaux et al. (2000) who introduce frontier-based estimates of plant capacity (see Johansen (1968)) as a foundation for this short-run industry model, thereby distinguishing between variations in technical efficiency and capacity utilization. This methodological refined model has been applied in analyzing excess capacities in fisheries and further extended in Kerstens et al. (2006). Starting from the ex-post fixity of investments in production capacities, this short-run Johansen (1972) model allows for some substitution possibilities by reallocating inputs and outputs among the units composing the industry while eliminating technical inefficiencies and major variations in capacity utilization among units. Furthermore, over time substitution and technical change can be traced via shifts in successive short-run industry models. None of the other above mentioned models accounts for the notion of production capacity or distinguishes clearly between technical inefficiency and variations in capacity utilization. As far as we know, this short-run industry model has never been applied to banking.

Since the goal of performance benchmarking in this case is prospective (i.e., providing management with strategic information to actually improve performance), there are strong reasons to believe that many people object to unobservable projection points implied by the traditional convexity hypothesis. This is evidenced in remarks, scattered in the literature, on the problems encountered in communicating the results of efficiency measurement to decision makers. We offer three examples. In a study applying convex nonparametric frontier methods to measure bank branch efficiency, Parkan (1987: 242) notes: "The comparison of a branch which was declared relatively efficient, to a hypothetical composite branch, did not allow for convincing practical arguments as to where the inefficiencies lay." In a similar vein, Bouhnik et al. (2001: 243), apart from criticizing extreme low scaling, also state: "… it is our experience that managers often question the meaning of convex combinations that involve what they perceive to be irrelevant DMUs." Finally, Epstein and Henderson (1989: 105) report similar experiences in that managers simply question the feasibility of the hypothetical projection points resulting from convex nonparametric frontiers. Thus, avoiding convexity may facilitate the implementation of frontier-based decision support models.[1] Therefore, in this contribution a lot of attention is devoted to testing for the impact of the convexity assumption in estimating capacity and in the results of the short-run industry model.

This contribution is structured as follows. We introduce in Section 2 the Johansen-Färe measure of plant capacity of the firm into a multi-output, frontier-based version of the short-run Johansen industry model. The first stage capacity model carefully checks for the impact of the convexity assumption on the estimated capacity utilization results. Policy scenarios considered for the short-run Johansen industry model vary in terms of their tolerance with respect to existing bank branch inefficiencies, the formulation of closure policies, the reallocation of labor in terms of integer units, etc. The data set of 142 bank branches of a German savings bank in the

[1] We thereby ignore the theoretical arguments against convexity based upon, for instance, the indivisibilities in production. See, e.g., Scarf (1994).

year 1998 is introduced in Section 3. The application to this German network of bank branches in Section 4 measures their efficiency and capacity utilization and demonstrate that by this industry model approach one can improve the performance of the whole branch network. A final section concludes and tries to outline some promising avenues for further research.

2. METHODOLOGY

2.1. Introduction

The theory of production is based on efficient technologies (production frontiers) and their value duals (such as minimal cost functions and maximum profit functions) and on envelope properties yielding cost-minimizing input demand functions and revenue maximizing output supply functions. In theory, emphasis is placed on efficient production and its consequences, and the evocative term "frontier" is applied to functions characterizing these boundaries. Using either parametric or nonparametric approaches, the standard cost structure is typically generated by imposing a specific functional form on the data and by obtaining the best fit by minimizing the deviations from the estimated structure. Efficiency measurement implies comparison between actual and optimal performance positioned on the relevant frontier. This frontier is called "best-practice", since it is an empirical approximation of the true but unknown frontier. The parametric approach is stochastically attempting to distinguish noise from inefficiency which requires strong assumptions, while the nonparametric approach does not run the risk of misspecification of the functional form but noise is not taken into account.[2]

We first offer several definitions to understand the mechanism of efficiency measurement. In general, efficiency analysis can be carried out at many levels of aggregation (i.e., at the plant, firm, industry or economy–wide level). The choice of level of aggregation is determined by –

[2] This is of course a simply presentation, but it presents the two essential differences between both approaches. For example, in recent years there has been a lot of work on the statistical foundation of the nonparametric approach: see Simar and Wilson (2008) for an overview.

among other things – availability of data and the purpose of the study. Here, we focus on the linkages between the efficiency both at the firm (branch) level and the industry (branch network) level. Economic efficiency has both a technical and allocative component. Technical efficiency is generally about avoiding waste, i.e., reducing the use of inputs given output levels or increasing outputs given input levels (see Koopmans (1951) for a formal definition). Allocative efficiency is referring to optimal proportions in outputs and inputs connected to prevailing relative prices.

When it comes to measurement of technical efficiency, the so-called Debreu (1951)-Farrell (1957) measure is used. In an output-augmenting orientation, the Debreu-Farrell measure is defined as the maximum radial expansion in all outputs that is feasible with given technology. From an engineering capacity concept, Johansen (1968) defined plant capacity as the maximal amount of output that can be produced per unit of time with an existing plant and its equipment without any restrictions on the availability of variable inputs. Capacity arises due to fixity of one or more inputs, and is thereby inherently a short-run concept. Färe (1984) formally showed the existence of plant capacity for certain types of production functions, while Färe et al. (1989) made the concept operational by using the Debreu-Farrell measure to calculate firm level capacity levels using nonparametric frontier approximations of technology. Their approach assumes that firms cannot exceed their use of fixed factors, but that their use of variable factors is unconstrained. A best-practice technology or frontier is constructed and the current output of each firm is evaluated against the maximum potential output at full capacity utilization, called "capacity output".

Summing these firm-level capacity outputs across firms offers an estimate of the aggregate industry capacity output. Comparing this aggregate industry capacity output to current industry output provides a measure of overcapacity at the industry level. However, neither firm-level technical measures nor firm-level capacity levels allow for reallocation of inputs and outputs across firms, precluding insight into the optimal restructuring and configuration of the industry. For example, the plant capacity measure implicitly assumes that production of capacity output is feasible and that the necessary variable inputs are available. In many other situations, relevant questions at the industry level are: What is the optimal firm-structure given the current aggregate output? How should the reallocation of inputs and outputs be performed between the firms? How does the reallocation look like if certain policy issues are taken into account? And what are the costs of pursuing these policy issues in terms of allocating more inputs than necessary?

To answer these questions, we combine the plant capacity notion (Johansen (1968)) at the individual and industry levels using a multiple-output and frontier-based version of the short-run Johansen (1972) sector model, a methodological refinement developed in Dervaux et al. (2000) and applied in, e.g., Kerstens et al. (2006). The short-run Johansen (1972) sector model analyses the industry structure resulting from underlying *ex post* firm-level production structures. Investment decisions imply a putty-clay production structure: while firms may eventually choose *ex ante* from a catalogue of production options exhibiting smooth substitution possibilities, most firms face fixed coefficients *ex post* and have a capacity that is entirely conditioned by the investment decision made. The short-run industry model nevertheless exhibits substitution possibilities when inputs and outputs can be reallocated across the units composing the industry. Over time, substitution and technical change can be traced via shifts in successive short-run industry models.

The revised short-run Johansen (1972) model proceeds in two phases. In a first step, the Johansen-Färe capacity measure determines capacity production for each individual firm at the production frontier. Second, this firm-level capacity information is employed in the industry model by a planning agency to select the level of activity at which individual firm capacities are utilized with the objective of minimizing fixed industry inputs given total outputs and capacities and the current state of technology. Following Dervaux et al. (2000) and Kerstens et al. (2006), the optimal industry or branch network configuration can be found by minimizing the total use of fixed inputs given that each firm cannot increase its use of fixed inputs and the production of the industry is at least at the current level.[3] The output level of each firm in this type of model is the capacity output estimated from the firm-level capacity model.

2.2. *Definitions of Efficiency, Plant Capacity, and the Short-Run Industry Model*

To develop these production models formally, the production technology S transforms inputs $x=(x_1,...,x_N)\in\mathbb{R}_+^N$ into outputs $u=(u_1,...,u_M)\in\mathbb{R}_+^M$ and summarizes the set of all feasible input and output vectors: $S=\left\{(x,u)\in\mathbb{R}_+^{N+M}:\ x \text{ can produce } u\right\}$. Let J be the number of firms/units ($j\in\left\{1,...,J\right\}$). The N-dimensional input vector x is partitioned into fixed factors (indexed by f) and variable factors (indexed by v): $x=(x_v,x_f)$. To determine the capacity output or technical efficiency, a radial output-oriented efficiency measure $E^0(x,u)=\max\left\{\theta:(x,\theta u)\in S\right\}$ is computed relative to a frontier technology providing the potential output given the current use of inputs, where restrictions on input use determine the precise nature of the measure.

Nonparametric inner-bound approximations of the true technology can be presented by the following set of production possibilities, assuming strong disposal of inputs and outputs and variable returns to scale (*VRS*):

[3] Remark that, when appropriate price information is available, the technical optimization (in terms of primal or quantity based aspects) in both stages of the short-run Johansen industry model can be replaced by alternative economic capacity notions in the first stage and economic objective functions (e.g., industry cost functions as in Førsund and Hjalmarsson (1983), or industry revenue or profit functions) in the second stage. In the first stage, economic capacity notions based on, e.g, the cost function can be employed (e.g., Prior (2003)).

$$S^{\Lambda,VRS} = \left\{ (x,u) \in \mathbb{R}_+^{N+M} : u_m \le \sum_{j=1}^{J} z_j u_{jm}, \quad m = 1,\ldots,M; \right.$$
$$\left. \sum_{j=1}^{J} z_j x_{jn} \le x_n, \quad n = 1,\ldots,N; \quad \sum_{j=1}^{J} z_j = 1, \quad z_j \in \Lambda, \quad j = 1,\ldots,J \right\}, \tag{1}$$

where $\Lambda \in \{C, NC\}$, with $C = \{z_j \in \mathbb{R}_+^J\}$ and $NC = \{z_j \in \mathbb{R}_+^J : z_j \in \{0,1\}\}$. $S^{\Lambda,VRS}$ assumes strong disposability of input and outputs, variable returns to scale, and it imposes either the traditional convexity (*C*) assumption or an alternative non-convexity (*NC*) hypothesis. From activity analysis, *z* is the vector of intensity or activity variables that indicates the intensity at which a particular activity is employed in constructing the reference technology by forming convex combinations of observations constituting the best practice-frontier.

From this same technology, a plant capacity version is defined by dropping the constraints on the variable input factors. This leads to Johansen's model definition of plant capacity whereby the availability of variable factors is unrestricted:

$$\hat{S}^{\Lambda,VRS} = \left\{ (x,u) \in \mathbb{R}_+^{N+M} : u_m \le \sum_{j=1}^{J} z_j u_{jm}, \quad m = 1,\ldots,M; \right.$$
$$\left. \sum_{j=1}^{J} z_j x_{jf} \le x_f, \quad f = 1,\ldots,F; \quad \sum_{j=1}^{J} z_j = 1, \quad z_j \in \Lambda, \quad j = 1,\ldots,J \right\}, \tag{2}$$

where Λ is again defined as above. To remain consistent with the plant capacity definition, in which only the fixed inputs are bounded at their observed level, the variable inputs in the production model (2) are allowed to vary at will to exploit the full capacity of outputs conditioned by the fixed inputs.

The efficiency measure θ_1 is found by solving the linear programming problem for each firm $j = 1, 2, \ldots, J$ relative to the production possibilities set with unrestricted variable inputs given by (2):

$$\max_{\theta_1^j, z_j} \left\{ \theta_1^j : (x, \theta_1^j u) \in \hat{S}^{\Lambda,VRS} \right\}. \tag{3}$$

The scalar θ_1 informs us by how much the production of each output of firm *j* can be increased. In particular, capacity output for firm *k* of the *m*th output is θ_1^{*k} multiplied by the actual production u_{km}. Hence, capacity utilization based on observed output (subscript 'oo') is:

$$CU_{oo}^{k} = \frac{1}{\theta_1^{*k}}. \quad (4)$$

Färe et al. (1994) note that this ray *CU* measure may be biased downwards, because there is no guarantee that the observed outputs are produced in a technically efficient way. The technical efficiency measure can be obtained by evaluating each firm $j = 1, 2, ..., J$ relative to the production possibility set $S^{\Lambda,VRS}$. The outcome (θ_2) shows by how much production can be increased using the given vector of inputs:

$$\max_{\theta_2^j, z_j} \left\{ \theta_2^j : (x, \theta_2^j u) \in S^{\Lambda,VRS} \right\}. \quad (5)$$

The technically efficient output vector is θ_2^{*k} multiplied by observed production for each output. Total industry output can be found by aggregating the firm-level technically efficient output $\theta_2^{*k} u_k$ of each firm. Likewise, the aggregate industry capacity output can be found as the sum of firm-level capacity outputs ($\theta_1^{*k} u_k$). The unbiased ray measure of capacity utilization given technically efficient output (subscript 'eo') is then:

$$CU_{eo}^{k} = \frac{\theta_2^{*k}}{\theta_1^{*k}}. \quad (6)$$

The focus here is on reallocation of resources between branches in a network by explicitly allowing improvements in technical efficiency and capacity utilization rates. The model is developed in two steps as follows. In the first step, from model (3), an optimal activity vector z^{*k} is provided for firm *k* and hence capacity output and its optimal use of fixed and variable inputs can be computed:

$$u_{km}^{*} = \sum_{j=1}^{J} z_j^{*k} u_{jm}; \quad x_{kf}^{*} = \sum_{j=1}^{J} z_j^{*k} x_{jf}; \quad x_{kv}^{*} = \sum_{j=1}^{J} z_j^{*k} x_{jv} . \tag{7}$$

In a second step, these "optimal" frontier figures (capacity output and capacity variable and fixed inputs) at the branch level are used as parameters in the industry model. In particular, the industry model minimizes the industry use of fixed inputs in a radial way such that the total production is at least at the current total level, or at a desired target level in the model extension developed below, by a reallocation of resources between firms or branches. Reallocation is allowed based on frontier production outputs and inputs used in each branch. In the short-run, it is assumed that current capacities cannot be exceeded either at the branch or industry level. Define U_m as the industry output level of output m and X_f (X_v) as the aggregate fixed (variable) inputs available to the sector of factor f (v):

$$U_m = \sum_{j=1}^{J} u_{jm}, \quad X_f = \sum_{j=1}^{J} x_{fj} \text{ and } X_v = \sum_{j=1}^{J} x_{vj} . \tag{8}$$

The formulation of the multi-output and frontier-based short-run Johansen (1972) industry model can then be specified as:

$$\begin{aligned}
&\min_{\theta, w, X_v} \theta \\
\text{s.t.} \quad &\sum_{j=1}^{J} u_{jm}^{*} w_j \geq U_m, \qquad m = 1,..,M, \\
&\sum_{j=1}^{J} x_{fj}^{*} w_j \leq \theta X_f, \qquad f = 1,...,F, \\
&\sum_{j=1}^{J} x_{vj}^{*} w_j \leq X_v, \qquad v = 1,...,V, \\
&0 \leq w_j \leq 1, \quad \theta \geq 0, \quad j = 1,...,J.
\end{aligned} \tag{9}$$

Rather than reflecting a returns-to-scale hypothesis, the variables w now indicate which firms' capacity is utilized and by how much. The components of the activity vector w are bounded above at unity, such that current capacities can never be exceeded. The first constraint prevents total production by a combination of firm capacities from falling below the current output levels. The second constraint means that the total use of fixed inputs (right-hand side) cannot be less than the use by a combination of firms. The third constraint calculates the resulting total use of variable inputs. Note that the total amount of variable inputs is a decision variable. The objective function is a radial input efficiency measure focusing on the fixed inputs solely. This input efficiency measure has a fixed-cost interpretation at the industry level. The activity vector *w* indicates which portions of the line segments representing the firm capacities are effectively used to produce outputs from given inputs.

To sum up, the optimal solution to this simple LP gives the combination of firms or branches that can produce the same or more outputs with less or the same use of fixed inputs in aggregate.[4] It measures the combined impact of the removal of any inefficiency, the exploitation of existing plant capacities, and the reallocation of inputs and outputs. Notice that an alternative could be to have an efficiency measure focusing on the expansion of industry outputs that has a revenue interpretation.

From a managerial point of view, the optimal solution of this short-run industry model provides information at two levels. First, at the level of the network it indicates the aggregate amount of variable inputs that is needed to realize the multiple aggregate outputs from given fixed aggregate inputs. If the optimal value of the aggregate variable inputs decision variable is larger than the current amount of aggregate variable inputs, then this implies additional recruitments are needed. Otherwise, a reduction in staff levels is required.

Second, at the level of the individual production units (bank branches) the model yields a complete planning for service production. Per unit, one obtains optimal fixed ($x^*_{fj} w^*_j$) and variable ($x^*_{vj} w^*_j$) inputs as well as optimal outputs ($u^*_{jm} w^*_j$). This may imply reallocations of inputs: fixed and variable inputs may be redistributed among units. Obviously, adjusting fixed inputs may be costly (e.g., renegotiating an existing office rental contract) and may furthermore

4 In fact, this short-run industry model is geometrically speaking a set consisting of a finite sum of line segments known as a zonotope (see Hildenbrand (1981: 1096)).

require time to implement (e.g., legal terms of notification prevent immediate changes). Equally so, adjusting variable inputs may be subject to a series of constraints (especially labor is under legal protection). This plan may also imply reallocations of outputs: this simply means that one adjusts the output targets within the planning horizon so as to better exploit the existing capacity of the whole network. Obviously, this may imply accompanying policy measures that are not necessarily part of the model (e.g., marginal changes in global and local marketing campaigns in an effort to gear consumer demand towards these targets).

2.3. Short-Run Industry Model: Additional Scenarios

Now, we turn to a discussion of some additional scenarios that extend the frontier-based short-run industry model to adapt to managerial concerns.

1. Restriction on number of branches:

Assume the number of branches should be restricted to N. Since the variable w_j represents the utilization of the corresponding branch, this restriction can be modeled with the following constraints:

$$\begin{aligned} & w_j \leq b_j \ (j=1,...,J); \\ & \sum_{j=1}^{J} b_j \leq N; \\ & b_j \in \{0,1\} \ (j=1,...,J). \end{aligned} \qquad (10)$$

By adding these constraints to model (9), it becomes a mixed integer program. The binary variable b_j indicates whether the corresponding branch is used in the optimal solution or not. The amount by which it is used can then be read from variable w_j.

2. Allow for existing inefficiency

The capacity outputs and the corresponding optimal fixed and variable inputs as computed in (7) presuppose that all eventually existing technical inefficiency can be eliminated in an effort to exploit the existing capacity of production. However, starting from the optimal activity vector $z^{*k} = (z_1^{*k}, ..., z_J^{*k})$ obtained from solving model (3), it is also possible to define capacity outputs and the corresponding optimal fixed and variable inputs while maintaining the existing levels of technical inefficiency by computing:

$$\bar{u}_{km}^{*} = \frac{1}{\theta_2^k}\sum_{j=1}^{J} z_j^{*k} u_{jm}; \quad x_{kf}^{*} = \sum_{j=1}^{J} z_j^{*k} x_{jf}; \quad x_{kv}^{*} = \sum_{j=1}^{J} z_j^{*k} x_{jv}. \qquad (11)$$

Hence, while the optimal fixed and variable inputs remain the same, the capacity outputs are maintained or scaled down by the measured amount of technical inefficiency (θ_2). Referring to the capacity output in (7) as the fully efficient one, the adjustment in (11) is called the fully inefficient capacity output. Both these capacity outputs can be considered special cases of the 100α % inefficient capacity output and the corresponding optimal fixed and variable inputs that can be defined as:

$$\bar{u}_{km}^{*}(\alpha) = \frac{1}{1+\alpha(\theta_2^k - 1)}\sum_{j=1}^{J} z_j^{*k} u_{jm}; \quad x_{kf}^{*} = \sum_{j=1}^{J} z_j^{*k} x_{jf}; \quad x_{kv}^{*} = \sum_{j=1}^{J} z_j^{*k} x_{jv}, \qquad (12)$$

with $0 \le \alpha \le 1$. Clearly, the 0% inefficient capacity output corresponds with the fully efficient capacity output, while the 100% inefficient capacity output coincides with the fully inefficient capacity output. When fully inefficient capacity output are used in the short-run industry model, this implies that one measures the impact of reallocation only.

3. *Restrictions on the personnel transfer*

Assuming the number of employees is a variable input, personnel transfer for a given branch with respect to the current situation is then measured by the difference between the optimal variable input resulting from the industry model and the observed variable input (i.e., $x_{vj}^{*} w_j - x_{vj}$). It could be meaningful to allow personnel transfer only in integer multiples of some unit β. For instance, $\beta = 0.5$ would mean that the number of employees must change in multiples of one half (e.g., because the basic unit of a labor contract in some countries is either a part-time of a full-time contract). Since this change can be either positive (reflecting an increase in number of employees) or negative (referring to a decrease), this condition can be modeled by the constraint:

$$x^*_{vj} w_j - x_{vj} = \beta(i_1 - i_2)\,, \qquad (13)$$

with i_1 and i_2 integer variables. The difference of both integer variables measures exactly the change in personnel expressed in units of β (e.g., $\beta = 0.5$ means this difference of integer variables measures personnel change in half units). Note that adding this type of constraint transforms model (9) to a mixed integer problem.

4. Imposing alternative aggregate output targets

If it is possible to impose alternative target values on the outputs, then the first set of constraints in model (9) needs to be changed to:

$$\sum_{j=1}^{J} u^*_{jm} w_j \geq (1+\gamma_m) U_m\,, \qquad (14)$$

with $\gamma_m \geq -1$. A value of $\gamma_m \geq 0$ (implying $1+\gamma_m \geq 1$) means that the aggregate output *m* of the industry model must be at least $100\gamma_m$% larger than the current industry level of output *m*. Obviously, positive values correspond with increases, while negative values reflect decreases with respect to the current industry level of output *m*. If all $\gamma_m = 0$, then no alternative target values are proposed and the original model (9) is obtained based upon observed aggregate outputs.

Remark that, in general, imposing a positive target value (i.e., above the output aggregate) additionally restricts the constraints. This leads to worse objective function values in the case of a minimization problem. Put differently, a positive target value leads to a higher efficiency measure θ. Ultimately, too large positive target values may result in infeasibilities.

By contrast, negative target values (i.e., below the output aggregate) relax the corresponding constraint, which results in a lower or equal efficiency measure value. Whether this phenomenon actually occurs, however, depends on the status of the corresponding constraint and on its relation with other constraints. For instance, adding a negative target value to a nonbinding output constraint has no influence on the optimal solution. Even if an output constraint is binding, other binding output constraints could prevent a reduction of the efficiency measure θ when adding a negative target value.

Additional scenarios that could eventually be envisioned are: (i) limiting the range of plant capacity utilization for the units in the optimal solution (see, e.g., Kerstens et al. (2006)), and (ii) aggregating some of the outputs to reduce the number of dimensions (at the risk that the required more spectacular changes are more difficult to implement).

3. DATA: BANK BRANCHES OF A GERMAN SAVINGS BANK

Data are obtained from the article by Porembski et al. (2005). These authors analyze a sample of 142 German bank branches in the year 1998. In this work, we measure the efficiency of these branches of a German savings bank and demonstrate that by a different industry model approach one can improve the efficiency over the whole network.

German thrift institutions are owned by communities or counties. Today, these institutions participate in all types of banking activities, either directly or through a central institution that is commonly owned. These banks are independent of each other, but share a number of resources. An important characteristic of these banks is that the goal of profit maximization is conditioned by the requirement of providing services to their stakeholders (e.g., community or county, to small businesses, and the middle-class). For example, nobody who wants to open an account can be rejected. These special characteristics cause some serious problems, since, for instance, it is not allowed to restrict branches to regions with profitable customer bases only. Moreover, increased competition is faced due to the globalization of financial markets, the spread of internet banking, and the increasing operational cost of personnel, whereas interest rates and profits have been decreasing over the last few years. This explains why these banks are very keen on increasing their productivity.

The bank analyzed is among the ten largest of its type in Germany. Its total assets in 1998 were in the tens of billions US $. To develop the bank branch industry model, we follow Porembski et al. (2005) and basically adopt a so-called, production approach to defining the transformation of banking inputs into financial services. Bank branches are considered as service providers to account holders performing transactions and processing documents. Outputs are therefore normally defined in terms of the numbers of transactions or documents processed. The outputs chosen cover most of the products offered by a branch and the level of disaggregation is high (e.g., one distinguishes between demand deposits for business and for households). However, very often, and also in this case, detailed transaction flow data are unavailable, whence the stock of the number of accounts of various types is employed instead. Furthermore, only physical inputs like labor and capital and their associated costs are taken into account. Actually, around 60% of the operating costs are due to personnel. Hence, the labor input is one of the most important at the branch level. A major part of the remaining operating costs are building and equipment costs. Since these costs are very difficult to determine (e.g., the corresponding book value is often biased), the input office space serves as a surrogate input measure.

Listing the inputs and outputs constituting the production technology in detail, the following inputs are available:

- Employees (number);
- Office space (square meters);

whereby the units of measurement are put in between braces. Notice that it is common to consider office space as a fixed input that cannot be modified in the short-run. Hence, employees are the sole variable inputs. In addition, there is information on the following 11 output dimensions:

- Private demand deposits (accounts);
- Business demand deposits (accounts);
- Time deposits (accounts);
- Saving deposits (accounts);
- Credits (accounts);
- Bearer securities (accounts);
- Recourse guarantees (accounts);
- Bonds (accounts);
- Investment deposits (accounts);
- Insurances (contracts);
- Contributions to a building society (contracts).

Descriptive statistics, including mean, variance, skewness, the minimum and the maximum, for these input and output dimensions are reported in Table 1. We can make the following observations. First, there is a lot of variation among these bank branches as witnessed by the standard deviation. Furthermore, the positive skewness of the distribution reveals the dominance of certain large units, mainly reflecting substantial differences in size. Second, notice that some branches do not seem to produce time deposits, recourse guarantees, or insurance since these outputs are zero at the minimum. This may reveal a variety of patterns of specialization among this sample bank branches. In addition, the last row contains the sum of all inputs and outputs at the level of the branch network. This serves as a benchmark to assess the impact of the various scenarios in the industry models.

< Table 1 about here >

4 EMPIRICAL RESULTS

First, we report extensively on the estimation results of the plant capacity measure and its underlying efficiency measures. We thereby focus on the impact of the convexity hypothesis and the impact of correcting the capacity definition for the presence of technical inefficiency or not. Thereafter, we turn to the basic results from the short-run industry model and also investigate the implied reallocations at the level of the individual branches. We thereby report on a series of different scenarios. Finally, we offer some examples of results at the individual branch level to show the level of detail at which these models provide guidelines.

4.1. Estimation of Plant Capacity: Testing for Convexity

Descriptive statistics for the capacity-related efficiency measure (θ_1), the ordinary technical efficiency measure (θ_2), and the plant capacity measure (CU_{eo}) are reported in Table 2 for both the convex and non-convex case. Four key observations can be made: (i) the output-oriented inefficiency measures are on average much higher in the convex case than in the non-convex case; (ii) in the non-convex case all bank branches except three are technically efficient in contrast to just about 40% of observations in the convex case; (iii) two thirds of all branches (97) operate at full capacity in the non-convex case compared to about one fifth (33) in the convex case; and (iv) these phenomena result in rather low average measures of capacity utilization in the convex case compared to the non-convex case.

< Table 2 about here >

The difference between the densities of the output efficiency measures obtained with the convex and non-convex models as well as the resulting ray CU measure can be tested with a statistic developed by Li (1996) and later refined by Fan and Ullah (1999). This test statistic has the critical advantage to be valid for dependent and independent variables, the former dependency being typical for frontier estimators. The null hypothesis states the equality of both distributions. Table 3 summarizes the obtained results. In total, three efficiency measures (θ_1, θ_2 and CU_{eo}), both in the convex and non-convex case, are compared two by two. Notice that the symmetry of the table immediately follows from the symmetry of the test itself. The values of these test statistics must be compared with the reference value for the target significance level. A value higher than the reference value leads to a rejection of the null hypothesis (implying that both density distributions can be considered statistically different). Table 3 also shows the conclusion depicted with symbols when tested for a significance level of 1%: an asterisk (*) is used when the null hypothesis is rejected (different densities) and an equality sign (=) flags that the null hypothesis cannot be rejected (equal densities). We notice that all density distributions can be considered different, except for θ_1 and CU_{eo} in the non-convex case. The latter exception is explained by the fact that only three observations are technically inefficient ($\theta_2 > 1$) in the non-convex case (hence, the ratio CU_{eo} is inevitably very close related to θ_1). In conclusion, statistical tests indicate that these efficiency measures follow different distributions. Put differently, adding the traditional convexity hypothesis is not as innocuous as it is traditionally assumed.

< Table 3 about here >

Table 4 reports descriptive statistics of plant capacity inputs and outputs for two variations: (i) convex vs. non-convex; and (ii) full efficiency vs. full inefficiency. These results need to be contrasted with the descriptive statistics on the inputs and outputs of the original data in Table 1. Comparing Tables 4 and 1, one immediately observes that: (i) the capacity inputs remain on average close to the observed inputs, while the choice for the output orientation of efficiency measurement implies that capacity outputs are quite above observed outputs; (ii) this divergence between capacity and observed outputs is more substantial for the convex case than for the non-convex case; and (iii) the difference between capacity outputs without and with technical inefficiency is again largest in the convex case. This analysis serves to underscore the importance of the convexity axiom and, to some lesser extent, the impact of eliminating technical inefficiency or not.

< Table 4 about here >

4.2. Short-Run Industry Model: Basic Results and Additional Scenarios

Instead of using the fully efficient capacity output in the short-run Johansen industry model formulated in (9), the fully inefficient capacity output (11) as well as the $100\alpha\%$ inefficient capacity output for a given α (12) can be employed, leading to a series of variations of this basic model. By examining these different models, the impact of allowing for inefficiency can be measured in combination with the difference between convex and non-convex estimates of capacity.

Table 5 summarizes exactly this impact of both convexity and inefficiency on several key decision variables. First, there is the influence on the optimal industry efficiency measure θ^*. In the next row, the influence on the number of branches is reported for which full capacity is used in realizing at least the aggregate outputs with only a fraction of the fixed aggregate inputs. Similarly, the next rows indicate the number of branches that are only partially used or not used at all to realize the set of constraints in model (9).

< Table 5 about here >

In the convex case, the effect of allowing for inefficiency is noticeable. We observe, for instance, an increase of the efficiency measure with 0.1 when allowing for all existing technical inefficiency (this is a relative increase of 17%). Since capacity outputs are lower when one allows for inefficiency, it is harder to economize on fixed inputs and an increase of its optimal value can indeed be expected. Furthermore, notice that the full efficiency case only utilizes 106 of the 142 branches. Since the number of branches only partially used is limited to only three, this means that 33 branches are not used at all to implement the optimal solutions obtained in the Johansen industry model. This is quite a substantial amount (23.2% of the total number of branches), making one doubt whether such solution is implementable in practice. When inefficiency is allowed for, then the number of unused branches is reduced to 28 (19.7%), which remains considerable.

Remark that, contrary to what one may expect, the branches that are no longer used in the optimal solution remain not necessarily the same when moving from the fully efficient to the fully inefficient case. Put differently, the 28 branches observed with zero capacity in the fully efficient scenario are not necessarily contained in the 33 branches that are no longer utilized in the fully efficient scenario. Examining the individual branches, we detect 11 of the 28 branches that are used in the fully efficient case but not used at all in the fully inefficient scenario. Except for one, these are even used at full capacity.

We end by looking at the results in the non-convex case. With respect to the optimal efficiency value θ^*, we notice only a minor increase of 0.003 (this is a relative increase of only 0.4%) when moving from the fully efficient to the fully inefficient industry model. From the individual results per branch, it can be observed that there is no shift in the optimal solution. Thus, all branches used at full capacity in the fully efficient case are also maintained at full capacity in the fully inefficient scenario. The same holds true for the branches used at partial capacity and for those that are no longer used at all. Only a minor change can be detected in the capacity of two branches used at partial capacity. Consequently, the effect of allowing inefficiency in the non-convex case can be neglected. The same holds for the other decision variables reported in this case, since there is no difference at all. Intermediate inefficiency levels for the non-convex model are therefore of limited interest in this particular study.

Notice that the number of unused branches reduces to 24 (16.9%) which is substantially lower compared to the convex model (33 in the fully efficient scenario and 28 in the fully inefficient case). From additional examination of individual branch results, it can be noticed that the 24 branches that are no longer used following the non-convex methodology are not necessarily contained in the unused branches according to the convex methodology. Indeed, with respect to full efficiency, 11 branches are found with zero capacity in the non-convex case, but with full capacity in the convex case. In the fully efficient scenario, even 13 branches can be detected having zero capacity according to the non-convex methodology, but with full capacity following the convex methodology. This underscores that the fundamentally different nature of the convex and non-convex technologies may have far reaching managerial consequences.

To complement Table 5, Figures 1a and 1b trace the evolution of the industry efficiency measure as a function of a given α for the convex and non-convex cases respectively. As could already be anticipated from considering the extreme cases in Table 5, the function for the convex case is much steeper because industry efficiency changes over a wider range. The relative flatness of this function in the non-convex case is related to the small amount of technical inefficiency that can be detected under this assumption in the first place.

< Figures 1a and 1b about here >

Notice that the industry efficiency measure has a fixed cost interpretation and denotes the potential budgetary gains from closing down the branches indicated by zero utilization in the industry model. However, one must realize that in practice a host of additional considerations may be necessary to choose among these in defining a coherent closure policy. As already pointed at previously, adjusting fixed inputs may be costly both when one is owner of the office space (e.g., should one rent out part of the excessive office space assuming this is technically feasible, or should one sell of the property and buy a smaller one somewhere nearby?) and when one is renting these (e.g., renegotiating an existing office rental contract may be costly). Furthermore, these changes require time to implement (e.g., legal terms in buying and selling contracts as well as in rental contracts prevent changes overnight). In addition, it may be necessary to include additional consideration into this decision making process. For instance, it makes a difference whether one closes down a branch in a town with two additional branches of the same bank or in a small village with no other branch around in the neighborhood. These decisions may thus need to be conditioned on a variety of geographical information that is currently ignored in the model.

We now restrict attention to the non-convex methodology. Furthermore, since the effect of allowing for inefficiency is negligible in the non-convex case, we also limit the analysis to the case of full efficiency. We discuss the following three scenarios of interest that have been formally introduced in subsection 2.3. Firstly, the impact of adding restrictions on the number of branches (10) in model (9) is considered. Secondly, we investigate the influence of adding restrictions on the personnel transfer (13) to the short-run industry model. Finally, we evaluate the effect of imposing some alternative aggregate output targets (see (14)). Results for all these scenarios are reported in Table 6.

< Table 6 about here >

Restrictions on the number of branches

The results of adding the constraints on the number of branches for some key reference values of N to the model are reported in the first five columns of Table 6. On one extreme, we notice that the problem becomes infeasible when limiting the number of branches to 95 or less. This means that we need at least 96 branches to deliver the current level of network outputs from given fixed inputs. On the other side of the range, we see that efficiency no longer improves when passing the limit of 118 branches. Furthermore, observe that in all cases, the number of branches used at full capacity is very close to the imposed limit N. Put differently, the number of branches used at partial capacity is very low (only one to two), meaning there seems to be little or no advantage of moving to scenarios that promote the use of partial capacities. Obviously, the value of the efficiency measure θ decreases as N increases. This observation corresponds with intuition since an increase in the number of branches implies using branches that are less efficient and/or that have less capacity.

Restrictions on the personnel transfer

Adding restrictions on the personnel transfer, the middle part of Table 6 reports the effect of adding such a restriction for two values of β. In particular, personnel transfer is only possible in integer multiples of either $\beta = 0.5$ (number of employees must change in multiples of one half) or $\beta = 1$ (number of employees must change in multiples of one). This scenario has two noticeable effects. First, the industry efficiency score increases substantially, implying that less fixed inputs can be economized. Second, there is a substantial move from branches working at full capacity to branches functioning at some partial capacity level. This actually turns out to be the only scenario producing such a result.

We add two remarks on potential implementation problems. First, the transfer of personnel can be difficult in view of geographical distances. For instance, it would make little sense to reallocate a person for say about 10% of his working time (about a half day per week in a five day working week) to a bank branch located at 500 km from his/her initial location. The current model ignores this issue basically because geographical information is lacking. However, in principle it is possible to extend the current model by restricting patterns of reallocation among units within a certain geographical radius (see, e.g., Giménez-García et al. (2007) for an example).

Second, the empirical model only employs aggregate information on personnel. Disaggregating personnel may yield more detailed results that are easier to implement and that have positive additional results. For instance, in Sherman and Ladino (1995) the efficiency results have been used to look for reductions in the number of branch managers by looking for possibilities to share managers for specific nearby bank branches. This again necessitates detailed geographical information. In a similar vein, the efficiency and capacity results could be used to make sure reallocations of managers go from high performance to low performance branches such that these relatively more successful managers can induce best practice behavior throughout the branch network.

Imposing alternative aggregate output targets

The last part of Table 6 reports on some aggregate output target scenarios. In a first scenario, we impose a positive output target of 10% on the number of saving deposits only. As a result, the optimal efficiency measure increases substantially from its original value of 0.702 to 0.775. To achieve this target, the number of branches needed at full capacity must be increased from 116 to 120, reducing the number of branches at zero capacity by 4. Increasing the target beyond 30% of current aggregate output is infeasible. For instance, using a negative reduction of 20% on the number of saving deposits has no influence at all on the optimal solution. Clearly, the other output constraints prevent such a reduction. When systematically looking for output variables that do have an influence when imposing a, for instance, 20% negative target, we observe that only the number of bearer securities accounts and the number of insurance contracts do make a difference. This effect is valid under ceteris paribus conditions, i.e., assuming no targets are imposed for the other outputs. First, in the case of the bearer securities, the efficiency measure θ is further reduced to 0.675, hereby using only 108 branches at full capacity compared to 116 originally (resulting in an increase of the number of unused branches from 24 to 32). Second, with respect to the number of insurance contracts, a more modest effect is observed: the efficiency measure only drops with 0.001. This result is obtained by utilizing 115 branches at full capacity instead of 116 initially. The number of branches no longer used remains the same (24), but when looking at individual results, we notice a minor shift. One branch previously not used is now used partially, and simultaneously another branch previously used only partially is now no longer used at all.

4.3. Results for Individual Bank Branches: Some Examples

In this subsection, we briefly report on the typical results at the level of individual observations generated from the above models. Table 7 contains results for a selection of individual bank branches. The columns contain the information for all input and output dimensions. The rows refer to specific computations and scenarios to be specified below.

< Table 7 about here >

We start by discussing the results for a technically inefficient unit operating below full plant capacity. In the first row one finds the data for the inputs and outputs of observation 71 that is representative for this case. First, in terms of technical efficiency, observation 71 yields an output efficiency measure of 1.544 and 1.096 for the convex and non-convex models respectively. Since plenty of studies explain the meaning of technical inefficiency (see, e.g., Färe et al. (1994)), the reader can consult these sources and we can safely ignore it here. Second, for the capacity model, the same observation obtains an output efficiency measure of 2.417 and 1.476 for the convex and non-convex models respectively. The resulting inputs and outputs at full capacity utilization are reported on rows 2 and 3 for the convex and non-convex models respectively. Unsurprisingly, capacity optimal inputs and outputs are largest in the convex case. For the non-convex model, observation 71 is compared to the single peer observation 103 (thus, row 3 simply represents the inputs and outputs of observation 103). For the convex model, the peers are a combination of several observations as can be identified by the optimal vector of activity variables (z^*). For the fixed input “Office space”, the peers use an equal (convex case) or a lower amount (non-convex case). For the variable input, by contrast, the peers use more of the variable input “Employees” (namely, 10.2 respect. 8 instead of 4) which is consistent with the plant capacity definition in which the variable inputs are allowed to vary at will. At full capacity, the outputs are larger than the ones produced by observation 71 and even substantially larger in the convex case.

In the industry model at full efficiency, observation 71 is maintained at full capacity ($w^* = 1$) in both the convex and non-convex models. This explains the numbers in rows 4 and 5, which exactly duplicate rows 2 and 3. In the industry model allowing for full inefficiency, the observation is maintained at full capacity ($w^* = 1$) in the non-convex model, but it is completely ignored in the convex solution ($w^* = 0$), which explains rows 6 and 7. Thus, bank branches apt for closing differ markedly between the different models.

Next, we comment upon the results for a bank branch that is technically inefficient under the convex model and technically efficient under non-convexity, while it does not operate at full plant capacity under either model. The first row in this second part of the table contains again the data for the inputs and outputs of observation 13, our example in this case. Ignoring the technical efficiency issue, the capacity model leads to an output efficiency measure of 2.745 and 1.349 for the convex and non-convex models respectively. For the latter model, observation 13 is compared to the peer observation 74. Corresponding inputs and outputs at full capacity utilization are again reported on rows 2 and 3. Similar remarks as to the difference between fixed and variable inputs apply once more.

In the industry model at full efficiency, observation 13 is maintained at full capacity ($w^* = 1$) in both the convex and non-convex models. When allowing for full inefficiency, the observation is again maintained at full capacity in the non-convex model and drops out of the convex solution. This explains rows 4 to 7.

Last but not least, we discuss briefly upon an observation that is efficient in both technical and capacity terms. In this case, the results for observation 4 can be summarized succinctly: it appears as it is in all industry model results. Thus, we can just list the observation itself on a single row to save some space.

5. CONCLUSIONS

Briefly summarizing the main contributions of this work, we focus shortly on the methodology employed as well as on the results. The efficiency literature analyzing the financial sector shows that even well performing banking institutions may have technical inefficiencies and some excess capacities at the level of their network of bank branches. Instead of relying on a burgeoning literature that starts from efficiency measurements at the individual level to come up with reallocations of resources at the firm level, we have opted to continue in the tradition of the revised short-run Johansen (1972) industry model, which is firmly grounded in the economics literature.

By way of example, we have analyzed the financial services supplied by a bank branch network of a rather large sized German savings bank (see Porembski et al. (2005)) using a production approach. The ordinary technical efficiency measure, the capacity-related efficiency measure, and the plant capacity measure have been computed using both convex and non-convex technologies. The resulting difference between the densities of these output efficiency measures and the resulting ray capacity utilization measure have been tested: the Li (1996) test statistic reveals that the resulting densities are almost all different from one another. This provides strong support to opt for a non-convex production technology rather than the traditional convex one for frontier benchmarking purposes.

Empirical results of the short-run industry model reveal a potential for closing down part of the network while maintaining current service levels, even under the most conservative estimates of efficiency and capacity (i.e., the ones based on a non-convex technology). Three additional scenarios related to the impact of adding restrictions on the number of branches on the one hand and on personnel transfer on the other hand, and the fixing of alternative aggregate output targets have also been documented.

Obviously, these scenarios do not exhaust the possibilities to adjust this network model to managerial needs. We have mentioned on several occasions the usefulness of including geographical information. Additional policy considerations could include local and regional market share considerations (competition issues in general). Obviously, while including these additional parameters need not be impossible, one must be aware that the inclusion of additional constraints lowers the potential benefits of the short-run industry model and that some combinations of constraints may even lead to infeasibilities.

The implementation cost of efficiency and capacity analysis and the resulting short-run industry models is high for single shot exercises, but this cost becomes low once the needed data on inputs and outputs are integrated into the accounting system (e.g., eventually as part of an activity based costing (ABC) strategy: see Kantor and Maital (1999)). Furthermore, while the computation of efficiency measures and capacity measures is rather straightforward and meanwhile a host of software options are around (e.g., in GAMS: see Olesen and Petersen (1996); in the freeware R: see Wilson (2008); in SAS: see Emrouznejad (2005), etc.), it is clear that the utilization of the short-run industry model as a strategic planning tool would ideally require its integration into a Decision Support System (DSS). We are unaware of written accounts reporting on the regular use of frontier benchmarking software in organizations.[5] This remains an important issue for future research.

Overall, we hope this contribution has shown convincingly that there is scope to employ efficiency-based models to manage bank branch networks both at a strategic and operational level. Obviously, more research is needed to come up with more detailed branch network models geared towards a more complete set of managerial needs.

[5] Non-convex frontier technologies have been used for years to assess credit union performance by their trade association (Credit Union National Association (CUNA)): see, e.g., Fried et al. (1995). CUNA recently launched "CU Benchmarker" as a web based, paid service for benchmarking to its member credit unions. See the CUNA webpage http://www.cuna.org/research/cu_benchmarker.html (consulted April 4, 2011).

Table 1: Descriptive Statistics of Inputs and Outputs

	Inputs		Outputs (all in numbers)										
	Personnel (number)	**Office space (m^2)**	**Private demand deposits**	**Business demand deposits**	**Time deposits**	**Saving deposits**	**Credit**	**Bearer securities**	**Recourse guarantees**	**Bonds**	**Investment deposits**	**Insurance**	**Contributions to a building society**
Mean	5.42	297.34	1846.91	272.31	37.32	5155.47	124.14	284.68	46.53	95.89	365.73	25.74	47.46
St. Dev.	4.17	213.12	1455.95	265.39	39.15	4086.80	100.01	196.27	44.45	85.79	288.29	26.67	48.81
Skew	1.58	1.71	1.68	2.19	2.78	1.78	1.56	1.48	1.98	2.07	1.79	2.67	2.22
Min.	1.0	64.00	432.00	31.00	0.00	1257.00	6.00	33.00	0.00	7.00	74.00	0.00	3.00
Max.	20.89	1228.00	7851.00	1563.00	285.00	20523.00	499.00	1020.00	271.00	503.00	1673.00	185.00	293.00
Total	769.84	42222	262262	38668	5300	732077	17628	40424	6607	13616	51934	3655	6739

Table 2: Descriptive Statistics for θ_1, θ_2 and CU_{eo}

	Convex			Non-Convex		
	θ_1	θ_2	CU_{eo}	θ_1	θ_2	CU_{eo}
Mean	1,533	1,147	0,801	1,086	1,002	0,939
St. Dev.	0,556	0,204	0,170	0,171	0,016	0,107
Min	1,000	1,000	0,343	1,000	1,000	0,565
Max	3,475	1,982	1,000	1,873	1,133	1,000
# Eff. Obs	33	57	32	97	139	97

Table 3: Li (1996) Test Statistic for Differences in Densities

		Convex			Non-Convex		
		θ_1	θ_2	CU_{eo}	θ_1	θ_2	CU_{eo}
Convex	θ_1	0.000 =	7.728 *	13.013 *	26.211 *	54.730 *	27.061 *
	θ_2	7.728 *	0.000 =	12.804 *	6.672 *	26.543 *	7.693 *
	CU_{eo}	13.013 *	12.804 *	0.000 =	27.074 *	53.955 *	24.841 *
Non-Convex	θ_1	26.211 *	6.672 *	27.074 *	0.000 =	6.205 *	0.506 =
	θ_2	54.730 *	26.543 *	53.955 *	6.205 *	0.000 =	6.215 *
	CU_{eo}	27.061 *	7.693 *	24.841 *	0.506 =	6.215 *	0.000 =

H_0: The two density distributions are equal. Conclusions: * : Reject H_0, = : Accept H_0.
Reference values: 1.28 for 10% sign. level, 1.64 for 5% sign. level, 2.33 for 1% sign. level.

Table 4: Descriptive Statistics of Plant Capacity Inputs and Outputs: Convex vs. Non-Convex; Full Efficiency vs. Full Inefficiency

	Personnel	Office space	Private demand deposits	Business demand deposits	Time deposits	Saving deposits	Credit	Bearer securities	Recourse guarantees	Bonds	Investment deposits	Insurance	Contributions to a building society
Full Efficiency	**Convex**												
Mean	8,13	296,37	2840,26	456,36	70,46	7642,41	212,73	398,15	86,95	155,05	562,67	42,14	71,86
St. Dev.	4,43	210,57	1524,33	306,12	47,83	4370,98	107,50	189,55	57,81	88,98	317,85	25,97	49,99
Min	2,00	64,00	552,00	46,00	0,00	1335,00	6,00	67,00	0,00	15,00	74,00	3,00	14,00
Max	20,89	1228,00	7851,00	1563,00	285,00	20523,00	499,00	1020,00	271,00	503,00	1673,00	185,00	293,00
	Non-Convex												
Mean	6,71	282,08	2308,20	365,72	54,79	6290,96	172,71	333,47	68,27	126,29	453,43	33,35	59,31
St. Dev.	4,95	204,79	1646,01	342,33	56,05	4695,95	124,40	191,35	71,28	98,43	329,60	29,33	52,76
Min	1,00	64,00	471,00	31,00	0,00	1335,00	6,00	57,00	0,00	14,00	74,00	0,00	3,00
Max	20,89	1228,00	7851,00	1563,00	285,00	20523,00	499,00	1020,00	271,00	503,00	1673,00	185,00	293,00

Full InefficiencyConvex											
Mean	2568,37	413,68	63,33	6922,23	191,76	358,87	77,81	140,60	509,14	38,37	65,72
St. Dev.	1542,24	305,14	46,58	4405,07	107,96	193,75	55,53	90,37	320,86	26,33	50,41
Min	552,00	46,00	0,00	1335,00	6,00	67,00	0,00	15,00	74,00	3,00	14,00
Max	7851,00	1563,00	285,00	20523,00	499,00	1020,00	271,00	503,00	1673,00	185,00	293,00
Non-Convex											
Mean	2303,87	365,13	54,68	6279,34	172,24	332,71	68,17	126,06	452,57	33,27	59,14
St. Dev.	1646,57	342,48	56,01	4696,89	124,14	191,27	71,31	98,45	329,67	29,32	52,67
Min	471,00	31,00	0,00	1335,00	6,00	57,00	0,00	14,00	74,00	0,00	3,00
Max	7851,00	1563,00	285,00	20523,00	499,00	1020,00	271,00	503,00	1673,00	185,00	293,00

Table 5: Basic Short-Run Industry Model Results: Impact of Convexity and Technical (In)efficiency

	Decision Variables	Full efficient ($\alpha = 0$)	Full inefficient ($\alpha = 1$)
Convex	Industry efficiency θ^*	0.588	0.688
	# Full Capacity w	106	112
	# Partial Capacity w	3	2
	# Zero Capacity w	33	28
Non Convex	Industry efficiency θ^*	0.702	0.705
	# Full Capacity w	116	116
	# Partial Capacity w	2	2
	# Zero Capacity w	24	24

Figure 1a: Industry Efficiency measure in Relation to α in Convex Case

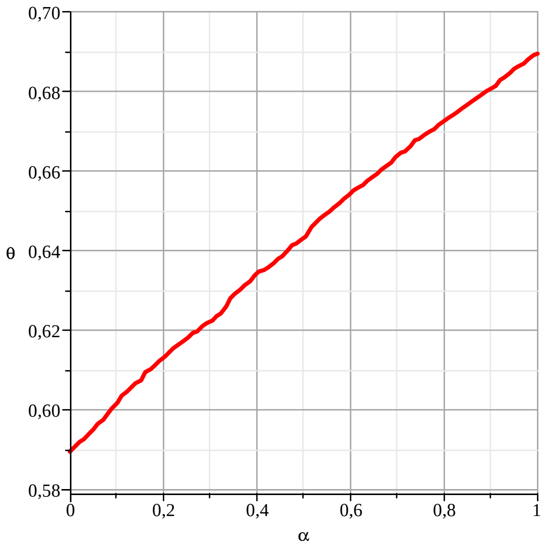


Figure 1b: Industry Efficiency measure in Relation to α in Non-convex Case

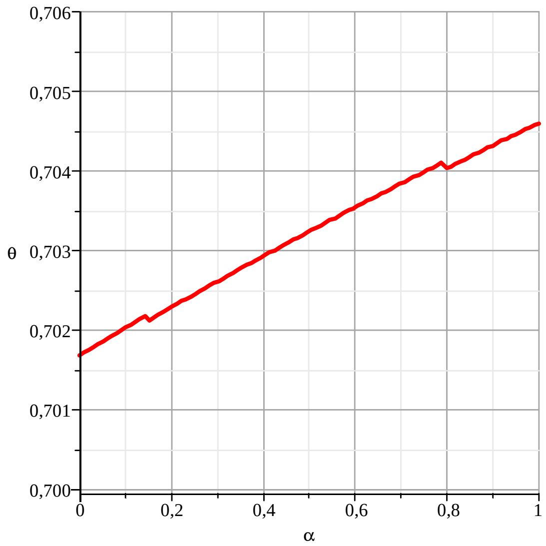


Table 6: Short-Run Industry Model Results: Additional Scenarios

	N					β		Aggregate output targets			
	≤ 95	96	100	117	≥ 118	0.5	1.0	S1*	S2	S3	S4
θ^*	–	0.766	0.722	0.702	0.702	0.711	0.723	0.775	0.702	0.675	0.701
# Full Cap.	–	95	99	115	116	84	78	120	116	108	115
# Partial Cap.	–	1	1	2	2	47	57	1	2	2	3
# Zero Cap.	–	46	42	25	24	11	7	21	24	32	24

* S1: Impose a target value of +10% on the number of saving deposits.
S2: Impose a target value of -20% on the number of saving deposits.
S3: Impose a target value of -20% on the number of bearer securities account.
S4: Impose a target value of -20% on the number of insurance contracts.

Table 7: Results for Individual Bank Branches: Some Examples

	Personnel	Office space	Private demand deposits	Business demand deposits	Time deposits	Saving deposits	Credit	Bearer securities	Recourse guarantees	Bonds	Investment deposits	Insurance	Contributions to a building society
Unit 71	4	310	1335	97	14	3821	72	191	23	52	324	17	21
Capacity C	10.20	310.00	3632.98	530.33	82.72	9234.53	273.67	461.61	128.65	175.51	783.04	41.09	66.42
Capacity NC	8	270	2914	352	108	8567	344	282	43	98	602	36	120
I. model C FE*	10.20	310.00	3632.98	530.33	82.72	9234.53	273.67	461.61	128.65	175.51	783.04	41.09	66.42
I. model NC FE	8	270	2914	352	108	8567	344	282	43	98	602	36	120
I. model C FI	0	0	0	0	0	0	0	0	0	0	0	0	0
I. Model NC FI	8.00	270.00	2658.01	321.08	98.51	7814.40	313.78	257.23	39.22	89.39	549.11	32.84	109.46
Unit 13	2	190	575	123	11	1614	32	54	10	41	186	7	14
Capacity C	8.24	190.00	2417.20	478.85	96.12	5726.16	169.82	249.46	143.47	130.74	510.64	35.74	38.44
Capacity NC	3.92	137	1537	166	27	3810	195	320	23	132	340	10	50
I. model C FE	8.24	190.00	2417.20	478.85	96.12	5726.16	169.82	249.46	143.47	130.74	510.64	35.74	38.44
I. model NC FE	3.92	137.00	1537.00	166.00	27.00	3810.00	195.00	320.00	23.00	132.00	340.00	10.00	50.00
I. model C FI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I. Model NC FI	3.92	137.00	1537.00	166.00	27.00	3810.00	195.00	320.00	23.00	132.00	340.00	10.00	50.00
Unit 4	2	64	813	57	9	2265	39	135	22	28	162	11	20

* FE = Full efficient / FI = Full inefficient

Chapter 2:

Malmquist and Hicks-Moorsteen Productivity Indices:

An Empirical Comparison Focusing on Infeasibilities

Abstract:

In the literature, two main ratio-based productivity indices have been defined in terms of a primal technology notion. While the Malmquist productivity index has become very popular, it has the problem that it is not always well-defined. By contrast, the less popular Hicks–Moorsteen productivity index is well-defined under weak conditions on technology. The purpose of this paper is to empirically illustrate the extent of this determinateness problem on some agricultural data sets under variations on the technology assumptions.

Keywords: Malmquist productivity index, Hicks-Moorsteen productivity index, Infeasibilities.

JEL classification: C43, D24, O33.

1. INTRODUCTION

Measuring multifactor or Total Factor Productivity (TFP) growth requires the construction of measures of input and output changes that incorporate changes in all dimensions. Two popular measures of TFP are the Törnqvist and the Fisher productivity indices (see, e.g., Coelli et al. (2005) or Färe, Grosskopf and Margaritis (2008)). Both use price information to aggregate the quantity data into quantity indices about the utilization of outputs and inputs. The resulting dual TFP index is a simple ratio of the thus obtained output and input quantity indices.

While dual TFP indices combine price and quantity information, a primal TFP index relies solely on quantity information summarizing the underlying production technology of the firm. The technology-based, discrete-time Malmquist productivity index introduced by Caves et al. (1982) constructs a production frontier representing technology and uses distance functions evaluated at different input–output combinations for productivity comparison.[1] In so doing, it offers a more general picture of productivity growth compared to other indices because: (i) the hypothesis of technical efficiency is relaxed; (ii) a decomposition is possible into technical efficiency changes and technology shifts (following Nishimizu and Page (1982), but see Zofío (2007) for further decompositions); and (iii) the computation of this index relative to multiple inputs and outputs technologies requires no price information (see Färe et al. (1995)). Meanwhile, the Malmquist productivity index has been widely applied in empirical research.

Another proposal for a discrete-time primal productivity index is the Hicks-Moorsteen or Malmquist TFP index, which is defined as a ratio of Malmquist output and input indices (see Bjurek (1996)). However, it is fair to say that it is less widely used in applied research

[1] Notice that Caves et al. (1982) in fact concentrated on showing how the traditional Törnqvist index approximates the technology-based Malmquist under certain conditions.

than the Malmquist productivity index (see, e.g., Bjurek et al. (1998) or Nemoto and Goto (2005)).

One well-known pitfall of the Malmquist index is that it is not always a TFP index (see Färe et al. (2008) for details). Indeed, rather quickly it was realized that its TFP properties are maintained under constant returns to scale, but as illustrated by Grifell-Tatjé and Lovell (1995) these are not preserved in the presence of variable returns to scale (i.e., a more general representation of technology). By contrast, Bjurek (1996) states that the Hicks-Moorsteen productivity index has a TFP interpretation and Grifell-Tatjé and Lovell (1999) illustrate this numerically. Later on, it has been proven that the Malmquist and Hicks–Moorsteen productivity indices coincide under two properties: (i) constant returns to scale, and (ii) inverse homotheticity (see Färe et al. (2008)). Thus, whenever these conditions are not satisfied in a sample or are not imposed on technology, both productivity indices diverge and the Malmquist index is a biased TFP measure (see O'Donnell (2008)). In the very few empirical applications we are aware of, both indices show a rather strong similarity, though they are not identical (see, e.g., Bjurek et al. (1998)).

Another problem known since the beginning of this literature is that some of the distance functions constituting the Malmquist productivity index may well be undefined when estimated using general technologies (see Färe et al. (1995)). However, empirical studies often ignore reporting on this infeasibility problem. For instance, Mukherjee, Ray and Miller (2001) report between 1% and 3.5% of infeasibilities per Malmquist index computed over a two-year period in a larger sample of 201 US commercial banks observed over several years. Briec and Kerstens (2009) prove that infeasibilities can occur for an even more general so-called Luenberger productivity indicator based upon more general directional distance functions. Thus, even this more general indicator does not satisfy the determinateness property in index theory.[2]

By contrast, the Hicks–Moorsteen index satisfies the determinateness axiom. This claim of Bjurek (1996) has been formally proven by Briec and Kerstens (2011) under mild conditions (i.e., mainly strong disposability of inputs and outputs).[3] This determinateness makes it a natural candidate to adopt it for benchmarking purposes focusing on the selection of a unit of strategic value against which performance is compared. Recently, Epure et al. (2011) define a series of variations based upon the Hicks–Moorsteen index offering such a benchmarking perspective.

O'Donnell (2008, 2010) has recently shown that profitability change can be decomposed into the product of a total factor productivity (TFP) index and an index measuring relative price changes. Many TFP indices can be decomposed into measures of technical change and technical efficiency change (following Nishimizu and Page (1982)), but furthermore into scale efficiency change and mix efficiency change components. Indices that can be decomposed in this way include the Fisher, Törnqvist and Hicks–Moorsteen TFP indices, but not the Malmquist index. Combined with the previously established doubts of the Malmquist index as a TFP index and the potential infeasibilities it may suffer from, one way to interpret it is as an index measuring local technical change and its components. We return to this interpretation in the concluding section.

While this infeasibility issue received rather limited attention in the productivity index literature, it is important when productivity indices are used for purposes of public policy. For example, the implementation of incentive regulatory mechanisms (e.g., in the context of price cap regulation) in a variety of network industries (gas, electricity, telecom, etc.) often makes

[2] Determinateness is one of Fisher's (1922) original axioms. It requires that an index remains well-defined even when one or more of its arguments become zero or infinity.

[3] Zaim (2004, 2006) employs a Hicks–Moorsteen index to measure environmental performance. He thereby imposes weak disposal in the bad outputs which are jointly produced along with the good outputs. Not entirely surprisingly, he reports some infeasibilities of this Hicks–Moorsteen environmental performance index.

use of multifactor best practice efficiency measures (see, e.g., Diewert and Nakamura (1999)). These regulatory applications would be seriously hampered when productivity change cannot be measured for some of the regulated firms (see, for example, Estache et al. (2007)).

Therefore, this contribution aims at empirically exploring the prevalence of the infeasibility problem and to document how the Hicks-Moorsteen index escapes from this problem. Section 2 provides the basic definitions of the technology, the various efficiency measures and distance functions, and the Malmquist and Hicks–Moorsteen productivity indices. Section 3 introduces the specifications of technologies used for computing these primal productivity indices. The next section describes the agricultural data sets employed. Section 5 discusses the empirical results. A final section concludes.

2. DEFINITIONS OF TECHNOLOGY AND PRODUCTIVITY INDICES

2.1 Technology and Distance Functions

The production technology uses the inputs $x=(x_1,...,x_n)\in\mathbb{R}_+^N$ to produce outputs $y=(y_1,...,y_m)\in\mathbb{R}_+^M$. In each time period ($t$) the set of all feasible input and output vectors is called the production possibility set ($T(t)$). This production possibility set is formally defined as follows:

$$T(t)=\left\{(x^t,y^t)\in\mathbb{R}_+^{N+M};\quad x^t \text{ can produce } y^t\right\}. \qquad (1)$$

This technology satisfies the following traditional assumptions: (T.1) no outputs without inputs; (T.2) infinite outputs are not allowed with a finite input vector; (T.3) closedness; and (T.4) strong input and output disposability. Note that the rather conventional convexity assumption is not always imposed (just when needed).

Debreu's (1951) coefficient of resources utilization and the efficiency measure introduced by Farrell (1957) are inversely related to the distance functions introduced by

Shephard (1970). In particular, the input-oriented efficiency measure of Debreu and Farrell $E^i_{T(t)}(x^t, y^t)$ is the inverse the Shephard (1970) input distance function. For the input-oriented case, this efficiency measure $E^i_{T(t)}(x^t, y^t)$ is based upon the minimum contraction of an input vector by a scalar λ to catch up with the boundary of technology:

$$E^i_{T(t)}(x^t, y^t) = \min_{\lambda}\left\{\lambda; \quad \left(\lambda x^t, y^t\right) \in T(t),\ \lambda \geq 0\right\}. \qquad (2)$$

In the case of an output efficiency measure, $E^o_{T(t)}(x^t, y^t)$ looks for the maximum expansion of an output vector by a scalar θ to catch up with the boundary of technology:

$$E^o_{T(t)}(x^t, y^t) = \max_{\theta}\left\{\theta; \quad \left(x^t, \theta y^t\right) \in T(t),\ \theta \geq 0\right\}. \qquad (3)$$

Notice that under constant returns to scale, already Førsund and Hjalmarsson (1979) have shown that input- and output-oriented efficiency measures are linked: $E^o_{T(t)}(x^t, y^t) = \left[E^i_{T(t)}(x^t, y^t)\right]^{-1}$.

For all $(a,b) \in \{t, t+1\} \times \{t, t+1\}$, the time-related version of the Farrell input- and output-oriented efficiency measure are given by

$$E^i_{T(a)}(x^b, y^b) = \min_{\lambda}\left\{\lambda; \quad \left(\lambda x^b, y^b\right) \in T(a),\ \lambda \geq 0\right\} \qquad (4)$$

and

$$E^o_{T(a)}(x^b, y^b) = \max_{\theta}\left\{\theta; \quad \left(x^b, \theta y^b\right) \in T(a),\ \theta \geq 0\right\}. \qquad (5)$$

Note that $E^i_{T(a)}(x^b, y^b) = +\infty$ if the set in (4) is empty and $E^o_{T(a)}(x^b, y^b) = -\infty$ if the set in (5) is empty.

2.2 Malmquist and Hicks–Moorsteen Productivity Indices

An input-oriented Malmquist productivity index $M^i((x^t, y^t),(x^{t+1}, y^{t+1}))$ can be defined as follows:

$$M^{i}((x^{t},y^{t}),(x^{t+1},y^{t+1})) = \left[\frac{E^{i}_{T(t)}\left(x^{t},y^{t}\right)}{E^{i}_{T(t)}\left(x^{t+1},y^{t+1}\right)} \frac{E^{i}_{T(t+1)}\left(x^{t},y^{t}\right)}{E^{i}_{T(t+1)}\left(x^{t+1},y^{t+1}\right)} \right]^{1/2}. \quad (6)$$

Its interpretation is as follows. Productivity growth (decline) is indicated by values smaller (larger) than unity. To avoid an arbitrary selection among base years, a geometric mean of period *t* (first ratio) and period *t+1* (second ratio) Malmquist indices is taken.[4]

A Hicks-Moorsteen (or Malmquist TFP) productivity index with base period *t* is defined as the ratio of a Malmquist output quantity index at base period *t* and a Malmquist input quantity index at base period *t*:

$$HM_{T(t)}((x^{t},y^{t}),(x^{t+1},y^{t+1})) = \frac{E^{o}_{T(t)}(x^{t},y^{t}) \big/ E^{o}_{T(t)}(x^{t},y^{t+1})}{E^{i}_{T(t)}(x^{t},y^{t}) \big/ E^{i}_{T(t)}(x^{t+1},y^{t})}. \quad (7)$$

When the Malmquist output quantity index (ratio in numerator) is larger (smaller) than unity, then more (less) outputs were produced in period t+1 than in period t from a given input vector. When the Malmquist input quantity index (ratio in denominator) is larger (smaller) than unity, then less (more) inputs were needed in period t+1 than in period t to produce a given output vector. When the Hicks-Moorsteen productivity index is larger (smaller) than unity, then it indicates productivity gain (loss).

In a similar way, a base period *t+1* Hicks-Moorsteen productivity index is defined as follows:

$$HM_{T(t+1)}((x^{t},y^{t}),(x^{t+1},y^{t+1})) = \frac{E^{o}_{T(t+1)}(x^{t+1},y^{t}) \big/ E^{o}_{T(t+1)}(x^{t+1},y^{t+1})}{E^{i}_{T(t+1)}(x^{t},y^{t+1}) \big/ E^{i}_{T(t+1)}(x^{t+1},y^{t+1})}. \quad (8)$$

Its interpretation is entirely similar to the above. A geometric mean of these two Hicks-Moorsteen productivity indices yields:

$$HM_{T(t),T(t+1)}((x^{t},y^{t}),(x^{t+1},y^{t+1})) = [HM_{T(t)}((x^{t},y^{t}),(x^{t+1},y^{t+1})) \cdot HM_{T(t+1)}((x^{t},y^{t}),(x^{t+1},y^{t+1}))]^{1/2}. \quad (9)$$

Again, its interpretation is entirely similar to the above.

[4] Notice that this geometric mean version does not chain. The choice for a fixed base version can readily remedy this issue.

2.3 Short-Run Malmquist and Hicks–Moorsteen Productivity Indices

Ouellette and Vierstraete (2004, 2010) define a short-run input-oriented Malmquist productivity index. To define a short-run Malmquist index, it is necessary to partition the input vector into a fixed and variable part ($x^t = (x^{f,t}, x^{v,t})$) such that there is always at least one variable input dimension.

It is now necessary to define a time-related version of the Farrell sub-vector input-oriented efficiency measure:

$$E^{f,i}_{T(a)}(x^b, y^b) = \min_{\lambda}\left\{\lambda;\quad \left(x^{f,b}, \lambda x^{v,b}, y^b\right) \in T(a),\ \lambda \geq 0\right\} \tag{10}$$

with $E^{f,i}_{T(a)}(x^{f,b}, x^{v,b}, y^b) = +\infty$ if the set in (10) is empty. The short-run input oriented Malmquist productivity index can now be defined as follows:

$$M^{f,i}((x^{f,t}, x^{v,t}, y^t),(x^{f,t+1}, x^{v,t+1}, y^{t+1})) = \left[\frac{E^{f,i}_{T(t)}(x^{f,t}, x^{v,t}, y^t)}{E^{f,i}_{T(t)}(x^{f,t+1}, x^{v,t+1}, y^{t+1})}\frac{E^{f,i}_{T(t+1)}(x^{f,t}, x^{v,t}, y^t)}{E^{f,i}_{T(t+1)}(x^{f,t+1}, x^{v,t+1}, y^{t+1})}\right]^{1/2}. \tag{11}$$

Its interpretation is similar to the input oriented Malmquist productivity index (6) defined above.

However, the above-mentioned sub-vector measure is sometimes undefined, i.e. it may not obtain a finite value. A base period *t* short-run Hicks–Moorsteen productivity index that is feasible can now be defined as follows:

$$HM^{f}_{T(t)}((x^t,y^t),(x^{t+1},y^{t+1})) = \frac{E^{o}_{T(t)}(x^t,y^t)\big/E^{o}_{T(t)}(x^t,y^{t+1})}{E^{i,f}_{T(t)}(x^{f,t},x^{v,t},y^t)\big/E^{i,f}_{T(t)}(x^{f,t},x^{v,t+1}y^t)}. \tag{12}$$

Its Malmquist output quantity index is identical to the one in the definition (7). But, its Malmquist input quantity index now focuses on reducing variable input dimensions only compared to fixed input and output dimensions that have the same time superscript as the technology (see (10) above). Analogously, a feasible base period *t+1* short-run Hicks–Moorsteen productivity index is defined as follows:

$$HM^{f}_{T(t+1)}((x^t,y^t),(x^{t+1},y^{t+1})) = \frac{E^{o}_{T(t+1)}(x^{t+1},y^t) \big/ E^{o}_{T(t+1)}(x^{t+1},y^{t+1})}{E^{i,f}_{T(t+1)}(x^{f,t+1},x^{v,t},y^{t+1}) \big/ E^{i,f}_{T(t+1)}(x^{f,t+1},x^{v,t+1}y^{t+1})}. \qquad (13)$$

Again, its Malmquist output quantity index is identical to the one in the definition (8), but the Malmquist input quantity index is constructed from the feasible Farrell sub-vector input-oriented efficiency measures in (10). A geometric mean of these two short-run Hicks-Moorsteen productivity indices is:

$$HM^{f}_{T(t),T(t+1)}((x^t,y^t),(x^{t+1},y^{t+1})) = [HM^{f}_{T(t)}((x^t,y^t),(x^{t+1},y^{t+1})) \cdot HM^{f}_{T(t+1)}((x^t,y^t),(x^{t+1},y^{t+1}))]^{1/2}. (14)$$

The interpretation remains entirely similar to the above indices.[5]

2.4. Infeasibilities in the Literature: A Selection

While it is not the purpose to give a systematic survey of all articles mentioning infeasibilities, we want to discuss a selection of articles to highlight some features of this phenomenon as documented in the current literature.

First, some authors are aware about the potential occurrence of infeasibilities in the specifications of technology they select to compute a Malmquist index, but they do not report any incidence about infeasibilities. For instance, Burgess and Wilson (1995) mention the possibility of infeasibilities in a VRS technology (page 350), but do not mention elsewhere any occurrence of infeasibilities affecting their bootstrap results. Similarly, Cummins and Rubio-Misas (2006) mention in the legend to their Tables 3 and 6 (page 346 and 352) the issue of infeasibilities, but give no indication about their eventual prevalence. It is not clear whether these authors indeed did not encounter any infeasibility or simply neglected reporting any cases.

Second, depending on the structure of the data the incidence of infeasibilities can range from extremely mild to extremely severe. For instance, Ray and Desli (1997) report just

[5] Note that this definition of a short-run Hicks–Moorsteen productivity index that is feasible is not trivial. See Briec and Kerstens (2011) for a variation on this short-run Hicks–Moorsteen index that is not well-defined.

one infeasibility for 1 out of 17 countries (5.8%) over the period 1979-1990 for two components of the Malmquist index (see their p. 1037). Bjurek et al. (1998) analyze an unbalanced panel of in total 2274 observations of Swedish electricity retail distributors over 21 years (1970-1990). This study reports on the productivity of three individual observations: for one of these, the first 10 out of 20 annual (50%) output-oriented Malmquist indices are infeasible (see Figure 5.5). There is no other indication in that contribution on infeasibilities in the sample. Finally, Silva Portela and Thanassoulis (2006) analyzed an unbalanced panel of Portuguese bank branches from March to December 2001 on a monthly basis. Starting with 57 branches and ending with 52 units due to branch closures, these authors report one month for which 47 out of 52 bank branches (90.4%) yield infeasible solutions.

Third, the incidence of these infeasibilities is to some extent conditioned by the assumptions imposed on the technology. Pastor et al. (2011) report up to 12% infeasibilities when using a VRS technology for a balanced panel of 93 firms over a 4 year period. Looking at the literature, it is striking that especially articles focusing on environmental performance using some combination of good and bad outputs, whereby the trade-off between both these outputs is modelled using a weak disposability assumption on the bad output, suffer from infeasible solutions. For instance, Yörük and Zaim (2005) analyze 28 countries over 16 years using four models with undesirable outputs. For some individual countries, these authors report up to 11 out of 16 (68.7%) infeasible solutions. Analyzing 41 countries over 22 years, Kumar (2006) states that three countries experience 2, 8 and 13 infeasible solutions, even when multiple year windows are used to mitigate the problem. Finally, in a study of 30 OECD countries from 2001 to 2002 Zhou and Ang (2008) report infeasibilities for 2 countries involved. The assumption of weak disposal being weaker than the traditional strong disposal assumption, one could conjecture that infeasibilities are somehow related to the volume of the technologies as resulting from the less or more demanding assumptions.

In the literature, it is often assumed that the imposition of constant returns to scale on technology guarantees feasibility of the Malmquist productivity index (see, e.g., Färe et al. (1994)).[6] However, to our knowledge no formal proof has ever been provided for this conjecture. Note that the feasibility of both the input-oriented and output-oriented Malmquist productivity index evaluated over all input respectively output dimensions can be inferred from results reported in Briec and Kerstens (2009) in a more general setting.[7] In the case of the input-oriented Malmquist productivity index, feasibility follows directly from Proposition 3.7 in Briec and Kerstens (2009). Assuming attainability is satisfied, it states that a necessary and sufficient condition for feasibility is that the underlying Farrell input efficiency measures aim for reductions over all input dimensions. For the case of the output-oriented Malmquist productivity index, Proposition 3.6 in the same article guarantees feasibility when the underlying Farrell output efficiency measures aim for reductions over all output dimensions. Note that both results are independent of the convexity axiom.

Some proposals in the literature somewhat change the discrete time nature of the Malmquist index by combining two or more years as a base technology. In particular, a global Malmquist index constructs a single global technology from all units and all time periods, while a biennal index just takes a time window of two (Pastor et al. (2011)) instead of several years (Kumar (2006)). Some of these proposals come at a serious drawback: e.g., the global Malmquist index requires re-computing all results whenever new time periods become available. The biennal index is attractive in that it always remains feasible, but in contrast to the fixed base period Malmquist index it is not transitive. The discussion of the relative merits of a fixed versus variable basis for productivity indices has been summarized in Färe et al.

[6] The seminal contribution of Färe et al. (1995) avoids this infeasibility problem in the output Malmquist productivity index by imposing a technology with a restrictive returns to scale assumption (in casu, non-increasing returns to scale).
[7] Indeed, Briec and Kerstens (2009) prove that infeasibilities may occur for the more general Luenberger productivity indicator based upon a directional distance function.

(2008). To avoid mixing up several other issues, we therefore ignore these proposals in our further discussion.

In conclusion, our analysis must be mainly seen in the light of this last evidence. We systematically explore the incidence of infeasibilities as a consequence of variations in the assumptions on technologies underlying the index computations.

3. METHODOLOGY: SPECIFICATION OF TECHNOLOGIES

In this section, we basically discuss the different specifications of technology relative to which both the Malmquist and the Hicks-Moorsteen productivity indices are computed. While parametric estimation of the component efficiency measures of the Malmquist index is possible (see, e.g., Atkinson et al. (2003) or Orea (2002)), the nonparametric estimation is far more popular.

We opt for nonparametric frontier technologies that impose neither an a priori functional form on technology, nor any restrictive assumptions regarding input remuneration. Moreover, the frontier nature of these technologies allows capturing any productivity inefficiencies and offers a benchmarking perspective.

Traditionally, most Malmquist productivity indices are computed relative to technologies imposing constant returns to scale. However, it is also possible to compute it with respect to a more flexible, variable returns to scale technology (see, e.g., Mukherjee et al. (2001)). In addition, it seems interesting to explicitly assess the impact of the traditional convexity hypothesis on the amount of infeasibilities in the Malmquist index. Even though convexity has been profoundly criticized (e.g., Briec et al. (2004)), rather few contributions have computed a Malmquist productivity index relative to a nonconvex technology (Tulkens and Malnero (1996) are among the exceptions), Let K be the number of firms/units ($k \in \{1,...,K\}$). Assuming strong disposal of inputs and outputs and some other maintained hypotheses to be specified below, following Briec et al. (2004) the nonparametric inner-bound approximations of the true technology can be presented by the following set of production possibilities:

$$S^{\Lambda,\Gamma} = \left\{ (x,y) \in \mathbb{R}_{+}^{N+M}; y_m \leq \delta \sum_{k=1}^{K} z_k y_{km}, \quad m=1,...,M; \right. \\ \left. \delta \sum_{k=1}^{K} z_k x_{kn} \leq x_n, \quad n=1,...,N; \quad z_k \in \Lambda; \quad \delta \in \Gamma; \quad k=1,...,K \right\}, \tag{15}$$

where $\Lambda \in \{C, NC\}$, with $C = \left\{ z_k \in \mathbb{R}_{+}; \sum_{k=1}^{K} z_k = 1 \right\}$ and $NC = \left\{ z_k \in \mathbb{R}_{+}; \sum_{k=1}^{K} z_k = 1, \quad z_k \in \{0,1\} \right\}$, and where $\Gamma \in \{CRS, VRS\}$, with $CRS = \{\delta; \delta \geq 0\}$ and $VRS = \{\delta; \delta = 1\}$.

From activity analysis, z is the vector of intensity or activity variables that indicates the intensity at which a particular activity is employed in constructing the reference technology by forming convex or nonconvex combinations of observations constituting the best practice frontier. This specification is nonlinear, but as shown by Briec et al (2004) it can be straightforwardly linearized in the convex cases (these methods are widely known under the moniker Data Envelopment Analysis (DEA)). However, the nonconvex cases involve solving either some nonlinear mixed integer programs, or some scaled vector dominance algorithms (these models are sometimes referred to as Free Disposal Hull (FDH)). More recently, Podinovski (2004) and Leleu (2006) have obtained mixed integer and linear programs respectively for all nonconvex specifications.

Thus, the calculation of each input-oriented Malmquist and the Hicks-Moorsteen productivity index requires efficiency measures that are computed according to four different approaches: convex (C) versus nonconvex (NC) technologies on the one hand, and constant

(CRS) versus variable returns to scale (VRS) on the other hand. Furthermore, some computations focus on a subset of variable input dimensions, while others do not distinguish between fixed and variable inputs. More detailed specifications of all efficiency measures involved in computing the input-oriented Malmquist and the Hicks-Moorsteen productivity indices are found in Appendix 1. The computations for the constant returns to scale case and reductions over all input dimensions on both convex and nonconvex technologies are just provided as a point of comparison.

4. SAMPLE DESCRIPTIONS

We employ two secondary agricultural data sets for our empirical analysis. The first sample is taken from Ivaldi et al. (1996) and contains an unbalanced panel of three years of French fruit producers. It is based on annual accounting data collected in a survey (Reseau d'Information Comptable Agricole (RICA)). Two criteria have been adapted in selecting farms: (i) the production of apples must be distinct from zero, and (ii) the productive acreage of the orchard must be larger than five acres. The technology combines three inputs to produce two outputs. The three aggregated inputs are: (i) capital (including land), (ii) labor, and (iii) materials. The sole fixed input dimension is capital. The two aggregate outputs are (i) production of apples, and (ii) an aggregate of other productions. While the years 1984, 1985 and 1986 count 130, 135 and 140 observations respectively, in the balanced panel only 92 observations are in common. This implies that 92 out of 184 observations or just 50% can be used in the index computations. Summary statistics and more details on the variable definitions are available in Appendix 2 in Ivaldi et al. (1996).

The second sample is a balanced panel of 43 smallholder rice farmers in the Tarlac region of the Philippines observed over the years 1990 to 1997 (see Coelli et al. (2005)). Fundamentally, four inputs are used to generate a single output (tons of freshly threshed rice).

The four inputs are: (i) area planted, (ii) labor used (in man-days of family and hired labor), (iii) fertilizer used (in kg of active ingredients), and (iv) other inputs used (a Laspeyres quantity index combining the inputs seed, herbicides, tractors and animals). The sole fixed input dimension is again area planted. Summary statistics and details on this sample are available in Appendix 2 of the book by Coelli et al. (2005).

5. EMPIRICAL RESULTS

First, we discuss the incidence of infeasibilities for the input-oriented Malmquist productivity index. Then, we discuss the descriptive statistics of both productivity indices as well as their rank correlations.

Table 1 offers descriptive statistics for the input-oriented Malmquist index as well as the Hicks–Moorsteen index. Part A of the table reports on the Ivaldi et al. (1996) data, while Part B contains the results for the Coelli et al. (2005) data. Each table is structured as follows. The upper part reports on the input-oriented Malmquist index, while the lower part contains the Hicks–Moorsteen index results. Within each part, the left-hand side of the table contains results for all input dimensions, while the right-hand side of the table focuses on a subvector of variable inputs. For each of these indices, the table distinguishes between convex and nonconvex on the one hand and between CRS and VRS results on the other hand.

< TABLE 1 ABOUT HERE >

The following tendencies are clearly apparent in terms of the incidence of infeasibilities for the input-oriented Malmquist productivity index. First, the amount of infeasibilities is greater under the VRS than under the CRS assumption. Second, the incidence of infeasibilities is larger under nonconvexity than under the traditional convexity hypothesis. Third, the number of infeasibilities is relatively higher under the subvector case compared to the standard index computed over all input dimensions. Notice that, by contrast, the Hicks-

Moorsteen productivity index always remains feasible (as could be expected from the Briec and Kerstens (2011) result).

These same tendencies can be illustrated using Figure 1, whereby models are indicated by combinations of the abbreviations for convex (C) versus nonconvex (NC), constant (CRS) versus variable (VRS) returns to scale, and the subvector case is denoted by "Subv". The percentage of infeasible observations in the input-oriented Malmquist index varies from 0% to 23.3% for the Ivaldi et al. (1996) data, and between 0% and 49.5% in the case of the Coelli et al. (2005) data. As conjectured above, these variations are clearly depending on the strength of the production axioms and the resulting volume of the technologies resulting from these less or more demanding assumptions.

< FIGURES 1 ABOUT HERE >

Obviously, as already illustrated in section 2.4, these percentages may reveal large variations across time and units. Therefore, Tables 2 and 3 report detailed descriptive statistics on infeasibilities over the years and the observations respectively. Both tables have a structure identical to Table 1 (except that the median is added and the last line is dropped). To facilitate comparisons, descriptive statistics are reported as percentages. To be explicit, for the statistics across periods in Table 2, we count the infeasibilities per comparison period over all units and we compute statistics across the available comparison periods. For the statistics across units in Table 3, we count the infeasibilities per observation over all time periods and we compute statistics across the units.

Starting with Table 2, recall that the Ivaldi et al. (1996) data only contain 3 years (hence 2 comparisons). While the standard Malmquist index shows between 3.26% and 5.43% infeasibilities across the board, the amount of infeasibilities in the subvector case varies considerably, with a clear peak for the nonconvex VRS case (up to 48.91% at maximum). Mean and median coincide and the standard deviation is rather small. For the

Coelli et al. (2005) data the standard Malmquist index shows a bit more variation, but the subvector case is even more seriously affected compared to the first sample case (the maximum now even goes up to 69.77%). Mean and median diverge and the standard deviation is now bigger.

Continuing with Table 3 (i.e., comparisons across units), one notices for the the Ivaldi et al. (1996) data that the standard Malmquist index has a lower mean than in Table 2.A, but a higher standard deviation and especially maximum (indeed several observations have no feasible solutions in the two comparison periods). For the Coelli et al. (2005) data the standard Malmquist index has at least an equal or higher mean than in Table 2.A, and a much higher standard deviation. Its maximum is again 100% (several observations have no feasible solutions in the seven comparison periods). In the the subvector case, the Ivaldi et al. (1996) data suffer on average less from infeasibilities, but its standard deviation and maxinmum is much higher. Exactly the same can be observed for the subvector case of the Coelli et al. (2005) data. Again, its maximum is 100%.

< TABLES 2 AND 3 ABOUT HERE >

To illustrate the large variations across years and units, we add some figures of some of the more extreme cases found in both data sets as follows. Figure 2 serves to illustrate the case of observation 18 in the Coelli et al. (2005) data that has 7 out of 7 infeasible solutions for the standard Malmquist index in case of a VRS assumption. Figure 2 plots the 1990 data represented by circles (o) and the 1991 data represented by crosses (x). The horizontal axis represents the first input variable (planted area), while the vertical axis shows the output (tons of freshly treshed rice). The black circle in the upper right position is DMU 18 in 1990. All frontier sections under all four technology assumptions are based on observations in 1991. When projecting DMU 18 in 1990 to the 1991 frontier in the input-orientation, there are feasible solutions for the CRS cases, while there are no feasible solutions for the VRS cases.

This infeasibility is due to observation 18 being situated above the VRS frontiers and the fact that an input-orientation has been adopted.

< FIGURE 2 ABOUT HERE >

The next Figure 3 illustrates the case of infeasibilities in the subvector Malmquist index. We have picked observation 35 of the Coelli et al. (2005) sample. From Figure 3.A, one notices two infeasibilities in the case of a convex VRS technology: one in the first and one in the last comparison periods. By contrast, in Figure 3.B four infeasibilities are observed in the case of a nonconvex VRS technology: one in the first and three in the comparison periods at the end. Each time, we contrast the subvector input Malmquist index with the Hicks-Moorsteen results, which do not suffer from this problem at all. Clearly, the infeasibility issue may complicate regulatory and managerial uses of the input-oriented Malmquist productivity index.

< FIGURES 3 ABOUT HERE >

Returning to the descriptive statistics in Table 1, we observe that on average the input-oriented Malmquist productivity index indicates a productivity decline, while the Hicks-Moorsteen productivity index clearly marks a productivity gain for both data sets. While the productivity change may seem strongest for the technology with the strongest assumptions with respect to returns to scale (i.e., CRS versus VRS) for the case of the input-oriented Malmquist index for the Ivaldi et al. (1996) data, this situation is exactly reversed for the Coelli et al. (2005) data. In a similar vein, while the productivity change may seem strongest for the technology with the strongest assumptions with respect to convexity (i.e., convexity versus nonconvexity) for the case of the input-oriented Malmquist index for the Ivaldi et al. (1996) data, this situation is again exactly reversed for the Coelli et al. (2005) data. While the subvector results are always slightly lower than the results on all input dimensions for the input-oriented Malmquist index for the Ivaldi et al. (1996) data, this situation is again exactly reversed for the Coelli et al. (2005) data. Similar observations can be made for the Hicks-Moorsteen productivity index. These observations can be explained as follows: while stronger (weaker) axioms lead to a larger (smaller) volume of the technology implying a clear ordering of efficiency measures in a static measurement, the productivity indices take ratios of efficiency measures in which such a clear ordering is simply absent.

Another, more relevant question is how the different productivity indices are correlated with one another. Table 4 reports rank correlation matrices between the different models for both data sets. First, the fundamental rank correlation between input Malmquist and Hicks-Moorsteen productivity indices is negative, since both indices move in opposite directions with respect to unity. Second, rank correlations between input Malmquist and Hicks-Moorsteen productivity indices are higher under constant returns to scale than under variable returns to scale, since constant returns to scale is among the necessary conditions for equality between both indices. Third, under variable returns to scale the rank correlations between input Malmquist and Hicks-Moorsteen productivity indices are quite high when other assumptions are identical, but sometimes rank correlations are even higher when one of the indices assumes constant returns to scale.

< TABLE 4 ABOUT HERE >

6. CONCLUSIONS

This contribution focuses on two discrete-time primal productivity indices that require a detailed knowledge of the underlying production technology: the Malmquist productivity index, and the Hicks-Moorsteen index. While the former has become immensely popular, the latter is still fairly little used in applied research. While the Hicks-Moorsteen productivity index has a TFP interpretation, it was already known that the Malmquist index is not always a TFP index and furthermore that the distance functions constituting it may well be undefined when estimated using general technologies. The recent work of O'Donnell (2010) casts further doubt on the Malmquist productivity index as a TFP index.

This contribution aims at empirically exploring the prevalence of the infeasibility problem for the popular Malmquist productivity index under a wide variety of assumptions on technology. It also has documented how the Hicks-Moorsteen index escapes from this problem. Apart from reviewing some of the results reported in the scant literature on this infeasibility issue, two agricultural data sets are employed to illustrate this issue in detail. It is found that the infeasibility problems seem directly linked to the strength of the assumptions defining the volume of the technology. Incidence varies from minor to very serious at the sample level. Especially at the level of individual observations the problem can be very severe: it can even inhibit any productivity evaluation in extreme cases.

What are the consequences of these empirical results for the interpretation of the Malmquist productivity index? These results clearly further undermine any remaining status of the Malmquist index as a TFP index. As mentioned before, one positive way out is to consider the Malmquist index as an index measuring local technical change (and eventually its components). In this perspective, the Malmquist productivity index attempts to provide a local answer (i.e., based on the observations observed in both time periods), but depending on the strength of the assumptions one is willing to impose on technology this answer is not always guaranteed. Thus, the local nature of its measurement comes at the cost of it not always being well-defined.

This new interpretation clearly makes the Malmquist productivity index much less suitable in a regulatory setting. We also expect more and more applied researchers to take an interest in the Hicks-Moorsteen index that suffers less from these problems and that has a clear TFP interpretation. It goes without saying that also other TFP indices may benefit from a renewed interest (see, e.g., the Fisher index in O'Donnell (2008) or in Ray and Mukherjee

(1996)). Another plausible consequence is that one may wonder whether it is meaningful to mix up these two structurally different types of productivity indices, as it has been done in certain methodological refinements.[8]

Finally, it is good to point out some limitations of this study. First, we compared the productivity indices, but ignored the eventual differences in their underlying decompositions (see Zofío (2007) for the Malmquist index and O'Donnell (2008, 2010) for the Hicks-Moorsteen index). Second, we have limited ourselves to productivity indices, but the same phenomena would most probably be observed for the pair of the Luenberger productivity indicator and its Luenberger-Hicks-Moorsteen counterpart (see Briec and Kerstens (2004)).

APPENDIX: COMPUTING MALMQUIST AND HICKS-MOORSTEEN INDICES: LINEAR PROGRAMMING PROBLEMS

For convenience, we explicitly specify all linear programming problems involved relative to a convex technology with constant returns to scale. For all other specifications, please consult Briec et al. (2004).

1. Malmquist productivity index: compute four radial efficiency measures per observation (x_0, y_0) under evaluation.

First within-period LP:

[8] For instance, some decompositions of the Hicks-Moorsteen productivity index (e.g., Nemoto and Goto (2005)) include components that are based on a Malmquist type of index and hence these could be infeasible, despite the fact that the overall index is well-defined. In a similar vein, some decompositions of the Malmquist productivity index (e.g., the input and output bias components in Färe, Grosskopf and Margaritis (2008): see their section 5.2.3) include components that are based on a Hicks-Moorsteen type of index. This situation is potentially confusing and probably requires some further reflection.

$$E^{i}_{T(t)}(x^{t}_{0}, y^{t}_{0}) = \min_{\lambda, z} \quad \lambda \tag{A.1}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} y^{t}_{km} z^{t}_{k} \geq y^{t}_{0m}, \qquad m = 1, \ldots, M,$$

$$\sum_{k=1}^{K} x^{t}_{kn} z^{t}_{k} \leq \lambda x^{t}_{0n}, \qquad n = 1, \ldots, N,$$

$$\lambda \geq 0, \ z^{t}_{k} \geq 0, \qquad k = 1, \ldots, K.$$

Second within-period LP:

$$E^{i}_{T(t+1)}(x^{t+1}_{0}, y^{t+1}_{0}) = \min_{\lambda, z} \quad \lambda \tag{A.2}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} y^{t+1}_{km} z^{t+1}_{k} \geq y^{t+1}_{0m}, \qquad m = 1, \ldots, M,$$

$$\sum_{k=1}^{K} x^{t+1}_{kn} z^{t+1}_{k} \leq \lambda x^{t+1}_{0n}, \qquad n = 1, \ldots, N,$$

$$\lambda \geq 0, \ z^{t+1}_{k} \geq 0, \qquad k = 1, \ldots, K.$$

First adjacent-period LP:

$$E^{i}_{T(t)}(x^{t+1}_{0}, y^{t+1}_{0}) = \min_{\lambda, z} \quad \lambda \tag{A.3}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} y^{t}_{km} z^{t}_{k} \geq y^{t+1}_{0m}, \qquad m = 1, \ldots, M,$$

$$\sum_{k=1}^{K} x^{t}_{kn} z^{t}_{k} \leq \lambda x^{t+1}_{0n}, \qquad n = 1, \ldots, N,$$

$$\lambda \geq 0, \ z^{t}_{k} \geq 0, \qquad k = 1, \ldots, K.$$

Second adjacent-period LP:

$$E^{t+1}_{i}(x^{t}_{0}, y^{t}_{0}) = \min_{\lambda, z} \quad \lambda \tag{A.4}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} y^{t+1}_{km} z^{t+1}_{k} \geq y^{t}_{0m}, \qquad m = 1, \ldots, M,$$

$$\sum_{k=1}^{K} x^{t+1}_{kn} z^{t+1}_{k} \leq \lambda x^{t}_{0n}, \qquad n = 1, \ldots, N,$$

$$\lambda \geq 0, \ z^{t+1}_{k} \geq 0, \qquad k = 1, \ldots, K.$$

The subvector cases require specifying separate sets of constraints for fixed and variable input dimensions and only radially reducing the variable input dimensions.

2. Hicks-Moorsteen productivity index: compute eight radial efficiency measures per observation (x_0, y_0) under evaluation.

While $E^i_{T(t)}(x^t_0, y^t_0)$ and $E^i_{T(t+1)}(x^{t+1}_0, y^{t+1}_0)$ are identical to the efficiency measures in the Malmquist index above, $E^o_{T(t)}(x^t_0, y^t_0)$ and $E^o_{T(t+1)}(x^{t+1}_0, y^{t+1}_0)$ are similar to the within same time period efficiency measures, only differing in measurement orientation.

The remaining four linear programs are:

$$E^o_{T(t)}(x^t_0, y^{t+1}_0) = \max_{\lambda, z} \quad \lambda$$
$$\text{s.t.} \quad \sum_{k=1}^{K} y^t_{km} z^t_k \geq \lambda y^{t+1}_{0m}, \quad m = 1,\dots,M,$$
$$\sum_{k=1}^{K} x^t_{kn} z^t_k \leq x^t_{0n}, \quad n = 1,\dots,N,$$
$$\lambda \geq 0, \; z^t_k \geq 0, \quad k = 1,\dots,K.$$

$$E^i_{T(t)}(x^{t+1}_0, y^t_0) = \min_{\lambda, z} \quad \lambda$$
$$\text{s.t.} \quad \sum_{k=1}^{K} y^t_{km} z^t_k \geq y^t_{0m}, \quad m = 1,\dots,M,$$
$$\sum_{k=1}^{K} x^t_{kn} z^t_k \leq \lambda x^{t+1}_{0n}, \quad n = 1,\dots,N,$$
$$\lambda \geq 0, \; z^t_k \geq 0, \quad k = 1,\dots,K.$$

$$E^o_{T(t+1)}(x^{t+1}_0, y^t_0) = \max_{\lambda, z} \quad \lambda$$
$$\text{s.t.} \quad \sum_{k=1}^{K} y^{t+1}_{km} z^{t+1}_k \geq \lambda y^t_{0m}, \quad m = 1,\dots,M,$$
$$\sum_{k=1}^{K} x^{t+1}_{kn} z^{t+1}_k \leq x^{t+1}_{0n}, \quad n = 1,\dots,N,$$
$$\lambda \geq 0, \; z^{t+1}_k \geq 0, \quad k = 1,\dots,K.$$

$$E^i_{T(t+1)}(x^t_0, y^{t+1}_0) = \min_{\lambda, z} \quad \lambda$$
$$\text{s.t.} \quad \sum_{k=1}^{K} y^{t+1}_{km} z^{t+1}_k \geq y^{t+1}_{0m}, \quad m = 1,\dots,M,$$
$$\sum_{k=1}^{K} x^{t+1}_{kn} z^{t+1}_k \leq \lambda x^t_{0n}, \quad n = 1,\dots,N,$$
$$\lambda \geq 0, \; z^{t+1}_k \geq 0, \quad k = 1,\dots,K.$$

The subvector cases require specifying separate sets of constraints for fixed and variable input dimensions and only radially reducing the variable input dimensions.

Table 1: Input-Oriented Malmquist vs. Hicks–Moorsteen Indices: Descriptive Statistics

Table 1.A. Ivaldi et al. (1996) Data

	Input-oriented Malmquist				Subvector Input-oriented Malmquist			
	Convex		Nonconvex		Convex		Nonconvex	
	CRS	VRS	CRS	VRS	CRS	VRS	CRS	VRS
Mean	1.1739	1.0321	1.1498	1.0631	1.1540	1.0269	1.1831	1.0244
Stand. Dev.	0.6747	0.3016	0.5852	0.3582	0.6985	0.3280	0.7334	0.4089
Min.	0.0854	0.5280	0.1305	0.4210	0.0606	0.4621	0.0989	0.3734
Max.	5.2365	2.5140	4.3777	2.8566	5.2365	2.7568	5.2506	3.1692
% Infeas. Obs.	0.00%	1.63%	0.00%	2.72%	1.90%	5.16%	4.62%	23.37%
	Hicks-Moorsteen				Subvector Hicks-Moorsteen			
	Convex		Nonconvex		Convex		Nonconvex	
	CRS	VRS	CRS	VRS	CRS	VRS	CRS	VRS
Mean	1.1502	1.1712	1.1157	1.1105	1.1881	1.1907	1.1602	1.1455
Stand. Dev.	1.0293	1.0082	0.7996	0.6664	1.0449	1.0129	0.8362	0.7142
Min.	0.1919	0.1925	0.2507	0.1668	0.1919	0.1925	0.2941	0.1668
Max.	11.6385	11.5746	7.0399	5.7377	11.6385	11.5746	7.6651	6.5508
% Infeas. Obs.	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 1.B. Coelli et al. (2005) Data

	Input-oriented Malmquist				Subvector Input-oriented Malmquist			
	Convex		Nonconvex		Convex		Nonconvex	
	CRS	VRS	CRS	VRS	CRS	VRS	CRS	VRS
Mean	1.0472	1.0639	1.0500	1.0726	1.0862	1.1190	1.0964	1.1497
Stand. Dev.	0.4484	0.4521	0.4465	0.4398	0.5978	0.6170	0.6721	0.5337
Min.	0.2659	0.1812	0.2790	0.3044	0.1648	0.2084	0.1619	0.3461
Max.	4.6811	3.2575	5.1048	3.2741	6.6767	6.4792	8.2680	3.9235
% Infeas. Obs.	0.00%	2.33%	0.00%	2.33%	3.65%	18.27%	3.65%	49.50%
	Hicks-Moorsteen				Subvector Hicks-Moorsteen			
	Convex		Nonconvex		Convex		Nonconvex	
	CRS	VRS	CRS	VRS	CRS	VRS	CRS	VRS
Mean	1.1024	1.0891	1.0966	1.0742	1.1077	1.0929	1.1048	1.0741
Stand. Dev.	0.4447	0.4335	0.4416	0.4156	0.4832	0.4661	0.4742	0.4549
Min.	0.2136	0.1729	0.1959	0.1664	0.2640	0.2508	0.2585	0.2576
Max.	3.7611	3.5994	3.5848	3.5483	4.4153	3.7866	4.1023	3.7126
% Infeas. Obs.	0.00%	0.00%	0.00%	0.00%	0.00%	0.33%	0.00%	0.00%

Figure 1: Input-Oriented Malmquist Index: Infeasibilities and Technology Assumptions

Figure 1.A. Ivaldi et al. (1996) Data

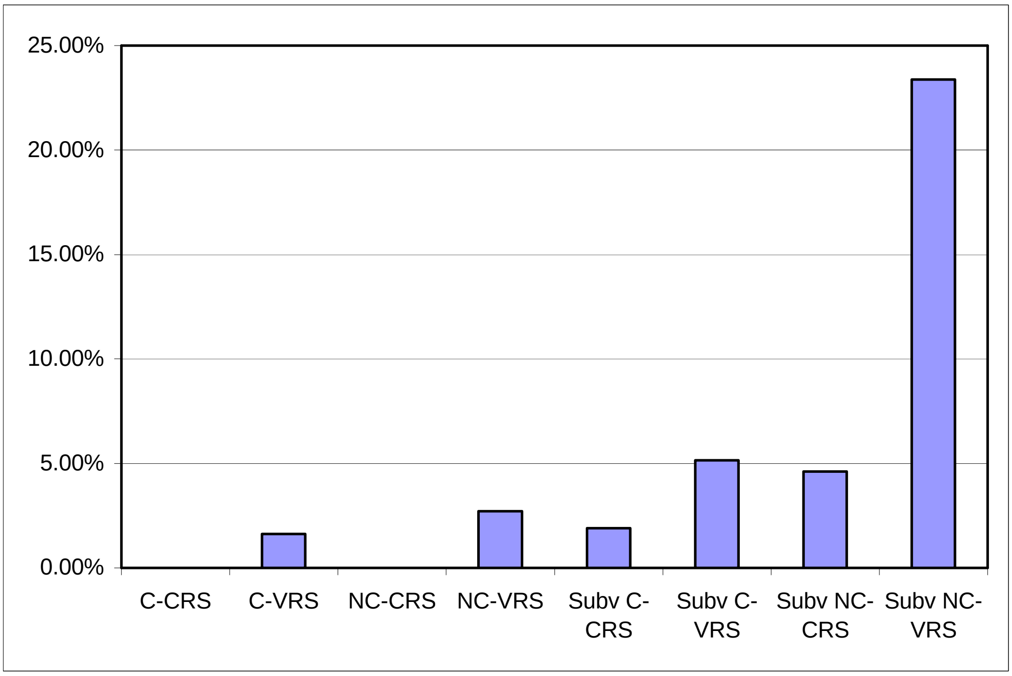


Figure 1.B. Coelli et al. (2005) Data

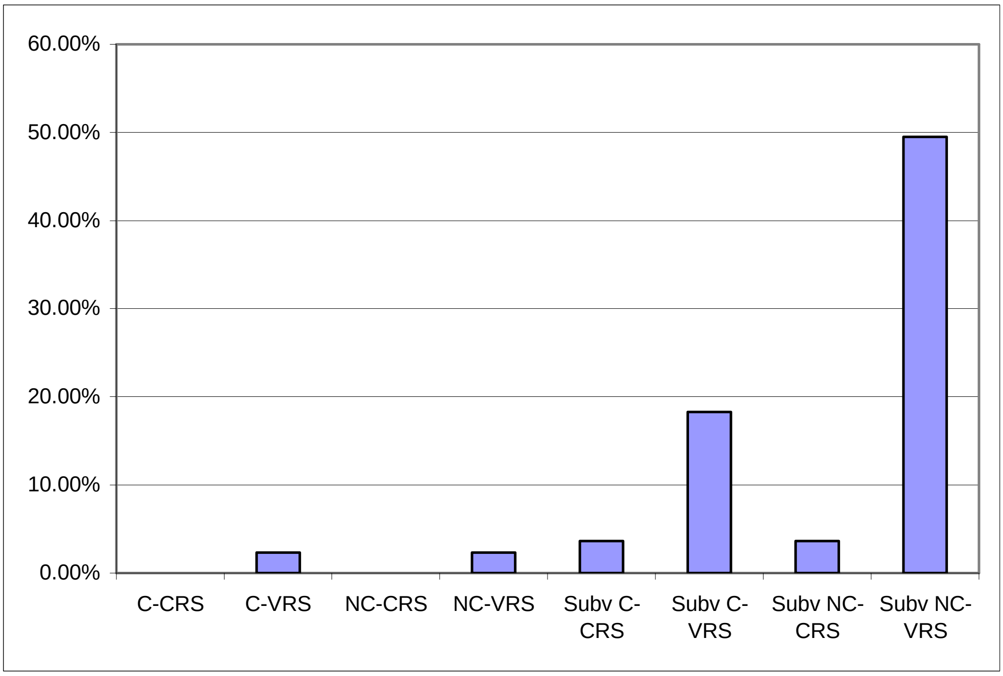


Table 2: Input-Oriented Malmquist: Descriptive Statistics of Infeasibilities Across Periods

Table 2.A. Ivaldi et al. (1996) Data (92 observations)

	Input-oriented Malmquist				Subvector Input-oriented Malmquist			
	Convex		Nonconvex		Convex		Nonconvex	
	CRS	VRS	CRS	VRS	CRS	VRS	CRS	VRS
Mean	0.00%	3.26%	0.00%	5.43%	3.80%	10.33%	9.24%	46.74%
Median	0.00%	3.26%	0.00%	5.43%	3.80%	10.33%	9.24%	46.74%
Stand. Dev.	0.00%	0.00%	0.00%	0.00%	0.77%	0.77%	0.77%	3.07%
Min.	0.00%	3.26%	0.00%	5.43%	3.26%	9.78%	8.70%	44.57%
Max.	0.00%	3.26%	0.00%	5.43%	4.35%	10.87%	9.78%	48.91%

Table 2.B. Coelli et al. (2005) Data (43 observations)

	Input-oriented Malmquist				Subvector Input-oriented Malmquist			
	Convex		Nonconvex		Convex		Nonconvex	
	CRS	VRS	CRS	VRS	CRS	VRS	CRS	VRS
Mean	0.00%	2.03%	0.00%	2.33%	3.65%	19.19%	3.65%	49.50%
Median	0.00%	2.33%	0.00%	2.33%	2.33%	11.63%	2.33%	46.51%
Stand. Dev.	0.00%	0.82%	0.00%	0.00%	1.83%	14.64%	1.83%	10.97%
Min.	0.00%	0.00%	0.00%	2.33%	2.33%	6.98%	2.33%	39.53%
Max.	0.00%	2.33%	0.00%	2.33%	6.98%	48.84%	6.98%	69.77%

Table 3: Input-Oriented Malmquist: Descriptive Statistics of Infeasibilities Across Units

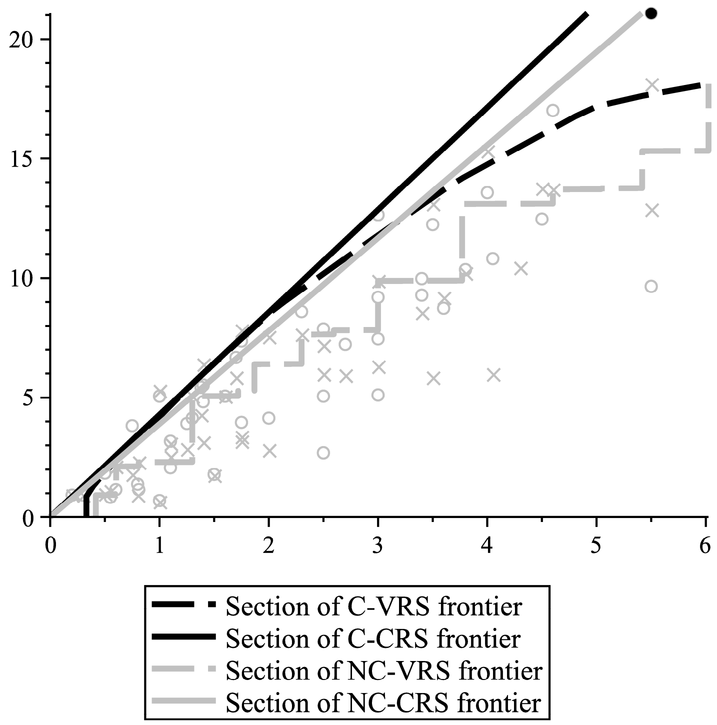
Table 3.A. Ivaldi et al. (1996) Data (2 time comparsions)

	Input-oriented Malmquist				Subvector Input-oriented Malmquist			
	Convex		Nonconvex		Convex		Nonconvex	
	CRS	VRS	CRS	VRS	CRS	VRS	CRS	VRS
Mean	0.00%	1.63%	0.00%	2.72%	1.90%	5.16%	4.62%	23.37%
Median	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Stand. Dev.	0.00%	12.70%	0.00%	15.44%	12.11%	19.22%	17.89%	38.73%
Min.	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Max.	0.00%	100.00%	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Table 3.B. Coelli et al. (2005) Data (7 time comparsions)

	Input-oriented Malmquist				Subvector Input-oriented Malmquist			
	Convex		Nonconvex		Convex		Nonconvex	
	CRS	VRS	CRS	VRS	CRS	VRS	CRS	VRS
Mean	0.00%	2.33%	0.00%	2.33%	3.65%	18.27%	3.65%	49.50%
Median	0.00%	0.00%	0.00%	0.00%	0.00%	14.29%	0.00%	57.14%
Stand. Dev.	0.00%	15.25%	0.00%	15.25%	12.51%	24.41%	12.51%	29.14%
Min.	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Max.	0.00%	100.00%	0.00%	100.00%	57.14%	100.00%	57.14%	100.00%

Figure 2: Input-Output Section With Position of Observation 18 (Coelli et al. (2005))



Legend: Observations are related to periods 1990-1991. Observations of 1990 are represented by circles (o). Observations of 1991 are represented by crosses (x). The horizontal axis represents the first input variable (planted area), while the vertical axis shows the output (tons of freshly treshed rice). The black dot in the upper right position is DMU 18 in 1990. All frontier sections are based on observations in 1991. When projecting DMU 18 in 1990 to the 1991 frontiers in the input-orientation, there are feasible solutions for the convex and nonconvex CRS cases, while there are no feasible solutions for the convex and nonconvex VRS cases.

Figure 3: Subvector Malmquist and Hicks-Moorsteen for Observation 35 (Coelli et al. (2005))

Figure 3.A. Convex VRS Technology

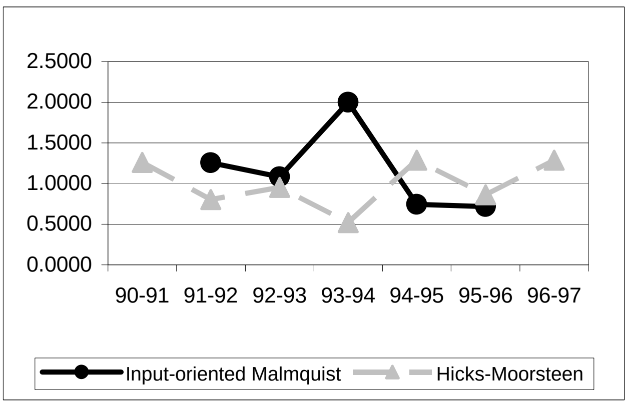


Figure 3.B. Nonconvex VRS Technology

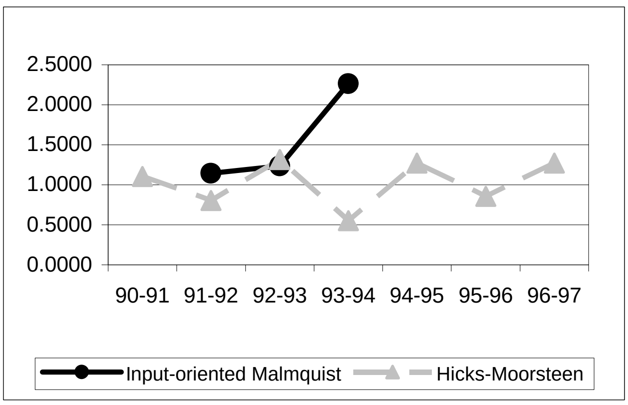


Table 4: Input-Oriented Malmquist vs. Hicks–Moorsteen Indices: Rank Correlations

Table 4.A. Ivaldi et al. (1996) Data: All Dimensions

			Input-oriented Malmquist				Hicks-Moorsteen			
			Convex		Nonconvex		Convex		Nonconvex	
			CRS	VRS	CRS	VRS	CRS	VRS	CRS	VRS
Input-	Convex	CRS	1.000							
oriented		VRS	0.674**	1.000						
Malmquist	Nonconvex	CRS	0.922**	0.590**	1.000					
		VRS	0.763**	0.705**	0.790**	1.000				
Hicks-	Convex	CRS	-0.999**	-0.672**	-0.923**	-0.760**	1.000			
Moorsteen		VRS	-0.945**	-0.706**	-0.909**	-0.784**	0.946**	1.000		
	Nonconvex	CRS	-0.909**	-0.559**	-0.984**	-0.772**	0.909**	0.882**	1.000	
		VRS	-0.887**	-0.580**	-0.910**	-0.785**	0.886**	0.914**	0.899**	1.000

*Note: Rank correlations were computed on between 87 and 92 available observations. ** Correlation is significant at the 0.01 level (2-tailed).*

Table 4.B. Ivaldi et al. (1996) Data: Subvector Case

			Input-oriented Malmquist				Hicks-Moorsteen			
			Convex		Nonconvex		Convex		Nonconvex	
			CRS	VRS	CRS	VRS	CRS	VRS	CRS	VRS
Input-	Convex	CRS	1.000							
oriented		VRS	0.675**	1.000						
Malmquist	Nonconvex	CRS	0.916**	0.595**	1.000					
		VRS	0.837**	0.658**	0.803**	1.000				
Hicks-	Convex	CRS	-0.999**	-0.676**	-0.919**	-0.825**	1.000			
Moorsteen		VRS	-0.942**	-0.696**	-0.927**	-0.806**	0.945**	1.000		
	Nonconvex	CRS	-0.925**	-0.594**	-0.960**	-0.792**	0.920**	0.913**	1.000	
		VRS	-0.860**	-0.596**	-0.883**	-0.794**	0.850**	0.907**	0.887**	1.000

*Note: Rank correlations were computed on between 51 and 92 available observations. ** Correlation is significant at the 0.01 level (2-tailed).*

Table 4.C. Coelli et al. (2005) Data: All Dimensions

			Input-oriented Malmquist				Hicks-Moorsteen			
			Convex		Nonconvex		Convex		Nonconvex	
			CRS	VRS	CRS	VRS	CRS	VRS	CRS	VRS
Input-oriented Malmquist	Convex	CRS	1.000							
		VRS	0.872**	1.000						
	Nonconvex	CRS	0.980**	0.848**	1.000					
		VRS	0.806**	0.870**	0.802**	1.000				
Hicks-Moorsteen	Convex	CRS	-1.000**	-0.872**	-0.980**	-0.806**	1.000			
		VRS	-0.951**	-0.884**	-0.932**	-0.821**	0.951**	1.000		
	Nonconvex	CRS	-0.980**	-0.848**	-1.000**	-0.802**	0.980**	0.932**	1.000	
		VRS	-0.873**	-0.798**	-0.880**	-0.810**	0.873**	0.903**	0.880**	1.000

*Note: Rank correlations were computed on between 294 and 301 available observations. ** Correlation is significant at the 0.01 level (2-tailed).*

Table 4.D. Coelli et al. (2005) Data: Subvector Case

			Input-oriented Malmquist				Hicks-Moorsteen			
			Convex		Nonconvex		Convex		Nonconvex	
			CRS	VRS	CRS	VRS	CRS	VRS	CRS	VRS
Input-oriented Malmquist	Convex	CRS	1.000							
		VRS	0.877**	1.000						
	Nonconvex	CRS	0.968**	0.839**	1.000					
		VRS	0.679**	0.780**	0.673**	1.000				
Hicks-Moorsteen	Convex	CRS	-0.974**	-0.864**	-0.926**	-0.674**	1.000			
		VRS	-0.921**	-0.854**	-0.870**	-0.678**	0.954**	1.000		
	Nonconvex	CRS	-0.936**	-0.837**	-0.913**	-0.674**	0.974**	0.944**	1.000	
		VRS	-0.802**	-0.691**	-0.766**	-0.701**	0.835**	0.868**	0.841**	1.000

*Note: Rank correlations were computed on between 151 and 301 available observations. ** Correlation is significant at the 0.01 level (2-tailed).*

Chapter 3:

Scale Economies and Returns to Scale

in Non-Parametric Models:

Exploring the Impact of Convexity

Abstract:

Returns to scale and economies of scale can be determined using production and cost functions. This contribution focuses on testing the empirical impact of convexity using non-parametric frontier specifications of technology and cost functions. Empirical results reveal the effect of convexity on estimates of scale efficiency and cost-based scale efficiency, as well as on the characterization of returns to scale for individual observations.

Keywords: Scale efficiency, Returns to scale; Economies to scale; Convexity.

JEL classification: D24.

1. INTRODUCTION

Efficiency and productivity analysis using frontier specifications of technology or value functions (e.g., cost functions) have become standard methods in the empirical toolbox of the applied researcher. These empirical studies serve a wide variety of academic, regulatory and managerial purposes. The traditional parametric and non-parametric approaches all maintain the axiom of convexity. However, indivisibilities in production imply that inputs and outputs are not completely divisible (see Scarf (1986, 1994)). Indivisibilities also put limitations on the up- or downscaling of the production processes. In addition, economies of scale and specialization may as well result in non-convex technologies (see, e.g., Romer (1990) on nonrival inputs in the new growth theory). In addition to the well-known case of externalities, all of these features of technology violate the convexity of the production possibility set (see Farrell (1959) for more details).

In some sectors the importance of non-convexities for cost determination has been clearly documented. For instance, non-convexities in electricity generation due to minimum up and down time constraints, multi-fuel effects, etc. leading to nonconvex and nondifferentiable variable costs have been documented in, e.g., Bjørndal and Jörnsten (2008) and Park et al. (2010). Furthermore, costs are non-convex in car manufacturing due to changes in the number of shifts and in the shutting down of plants for some time (see Copeland and Hall (2011)). It is widely acknowledged that many operations management problems in industry and distribution involve some form of indivisibilities requiring integer optimisation. However, most economic literature ignores such non-convexities in production.

In the non-parametric approach to production theory, the Free Disposal Hull (FDH) model -introduced by Deprins et al. (1984)- was originally designed to relax the convexity assumption underlying the traditional convex models (known under the moniker Data Envelopment Analysis (DEA)), whereby free disposal of inputs and outputs becomes the key assumption. A step towards extending the potential of FDH was initiated in Kerstens and Vanden Eeckaut (1999). These authors introduce specific returns to scale assumptions into the basic FDH model and propose a new goodness-of-fit method to infer the characterization of returns to scale for non-convex technologies.[14] Another step extending the scope for non-convex production modeling is found in Briec et al. (2004). These authors propose non-convex cost functions that are always larger or equal to their convex counterparts and they relate the traditional convex decomposition into technical, scale, allocative and overall efficiency with its non-convex counterparts.

The aim of this contribution is to explore the differences between technical and scale efficiencies based on both traditional convex and these rather new non-convex technology and cost function estimations. Such estimates have never been reported based on the cost function. More importantly, we also illustrate the eventual differences between the characterization of economies of scale and returns to scale for convex as well as non-convex cost functions and technologies. This has –to the best of our knowledge- never been reported in an international publication.

If differences in technical and scale efficiencies as well as the characterization of economies of scale and returns to scale for individual observations turns out to be conditioned by the convexity assumption, then this has important consequences. For one, investment decisions to increase or decrease the scale of operations based on economies of scale or returns to scale information could be responding to the wrong signals. For another, some capacity notions are very closely linked with the notion of scale economies. For instance, economic capacity as defined by Cassel (1937) and Klein

[14] This method was further refined for the case of convex technologies in Briec et al. (2000). Podinovski (2004a) extended this approach by introducing a distinction between local and global returns to scale.

(1960), among others, considers the outputs determined by the minimum of the long run average total costs as a reference to determine practical capacity utilisation ratios.

The next section defines the technology and the cost function and introduces some basic efficiency decompositions. Section 3 introduces the specific production and cost models to be estimated and develops the method to characterize returns to scale for individual observations in a convex and a non-convex setting. Next, we introduce the data sets employed in the empirical application. Section 5 presents the empirical results in detail. A final section concludes.

2. TECHNOLOGY, COST FUNCTION AND EFFICIENCY DECOMPOSITION

2.1. Definitions of Technology and Cost Functions

We start by defining technology and some basic notation. Denoting an n-dimensional input vector ($x \in \mathbb{R}^n_+$) and an m-dimensional output vector ($y \in \mathbb{R}^m_+$), the production possibility set or technology is defined as follows: $T = \{(x,y) \mid x \text{ can produce } y\}$. The input set associated with T denotes all input vectors x capable of producing a given output vector y: $L(y) = \{x \mid (x,y) \in T\}$. This input set $L(y)$ associated with T satisfies some combination of the following standard assumptions: apart from the traditional regularity conditions (i.e., no free lunch and the possibility of inaction, boundedness, closedness, and strong disposal), the key assumptions of importance for our contribution are convexity or not of the input set, and constant returns to scale (see, e.g., Färe et al. (1994) for details).

Since we only treat part of the static efficiency decomposition in the input orientation, we first define the radial input efficiency measure as:

$$DF_i(x, y) = \min\{\lambda \mid \lambda \geq 0, (\lambda\, x) \in L(y)\}. \quad (1)$$

This measure is simply the inverse of the input distance function and therefore offers a complete characterization of technology. Its properties are discussed in great detail in, for instance, Färe et al. (1994). Most importantly: (i) $0 < DF_i(x,y) \leq 1$, with efficient production on the boundary (isoquant) of $L(y)$ represented by unity; (ii) it has a cost interpretation.

Turning to a dual representation of technology, the cost function is defined as the minimum expenditures to produce an output vector y given a vector of semi-positive input prices ($w \in \mathbb{R}^n_+$):

$$C(y,w) = \min\{wx \mid x \in L(y)\}. \quad (2)$$

The fact that non-convexity has an impact on the duality relations between distance functions with cost, revenue and profit functions is widely ignored. For instance, costs evaluated on non-convex technologies are clearly higher or equal to costs evaluated on convex technologies (see Briec et al (2004)). This relation simply reflects the property that cost functions are non-decreasing in outputs and convex (non-convex) in the outputs depending on whether the technology is convex (non-convex) (see Jacobsen (1970): Proposition 5.2). The same reasoning applies to the revenue function and all variations of the profit function, except of course the long run profit function where convexity indeed does not matter for duality. Advanced micro-economic textbooks ignore this issue when, for instance, discussing the properties of the cost function (see Jehle and Reny (2000: p. 129) or Mass-Collell et al. (1995: p. 141)).

2.2. Basic Efficiency Decompositions

Farrell (1957) was the first to distinguish between technical and allocative efficiency in a frontier context. In addition, Banker et al. (1984), Färe et al. (1983), and

Førsund and Hjalmarsson (1974, 1979) all distinguish between technical and scale efficiency exploiting the distinction between technologies with constant (*CRS*) and variable (*VRS*) returns to scale. Since technologies vary, amongst others, in terms of underlying assumptions regarding returns to scale, the above notation of the efficiency measure and cost function can be conditioned on the difference between constant and variable returns to scale (convention: *C*=*CRS*, *V*=*VRS*). In this subsection we also explore the impact of convexity on the efficiency components of convex respectively non-convex decompositions.

Definition 1: *Under the above assumptions on the input set L(y), the following input-oriented efficiency notions can be distinguished:*

1) *Technical Efficiency is the quantity:* $TE_i(x,y) = DF_i(x,y \mid V)$.

2) *Overall Technical Efficiency is the quantity:* $OTE_i(x,y) = DF_i(x,y \mid C)$.

3) *Scale Efficiency is the quantity:* $SCE_i(x,y) = DF_i(x,y \mid C)/DF_i(x,y \mid V)$.

4) *Economic Efficiency for given scale is the quantity:* $OE_i(x,y,w \mid V) = C(y,w \mid V)/wx$.

5) *Overall Economic Efficiency is the quantity:* $OE_i(x,y,w \mid C) = C(y,w \mid C)/wx$.

6) *Cost-based Scale Efficiency is the quantity:*

$$CSCE_i(x,y,w) = \frac{C(y,w|C)/wx}{C(y,w|V)/wx} = \frac{OE_i(x,y,w|C)}{OE_i(x,y,w|V)}.$$

First, technical efficiency ($TE_i(x,y)$) presupposes that production occurs at the boundary of a VRS technology. Otherwise, a producer is technically inefficient. $TE_i(x,y)$ is traditionally evaluated relative to a *VRS* technology with strong disposability using $DF_i(x,y \mid V)$. Second, overall technical efficiency ($OTE_i(x,y)$) requires that production occurs on the boundary of a *CRS* technology. Otherwise, it is overall technically inefficient. Third, scale efficiency ($SCE_i(x,y)$) implies that the choice of inputs and outputs is optimal from the viewpoint of a long run *CRS* technology. Otherwise, a producer is scale inefficient. $SCE_i(x,y)$ results from comparing an observation to *CRS* and *VRS* technologies while maintaining strong disposability.

Turning to the dual variation on the same decomposition, one distinguishes between "economic efficiency given the scale" and "economic efficiency" (in the terminology of Seitz (1970)) depending on whether cost efficiency is measured relative to a VRS or CRS technology. Overall efficiency relative to VRS ($OE_i(x,y,w|V)$) requires estimating a total cost function relative to a *VRS* technology and taking a ratio of these minimal to actual costs. If the optimal input mix has in fact been chosen, then the organisation is overall efficient; otherwise, it is overall inefficient. Overall efficiency relative to CRS ($OE_i(x,y,w|C)$) requires estimating a total cost function relative to a *CRS* technology and otherwise it has a similar interpretation. Finally, following Seitz (1970, 1971), scale efficiency can also be based on a dual characterisation of technology by comparing overall efficiency measures defined with respect to different technologies. Cost-based scale efficiency $CSCE_i(x,y,w)$ requires that the choice of inputs and outputs is optimal from the viewpoint of an ideal long run *CRS* cost function. Otherwise, a producer is cost-based scale inefficient. $CSCE_i(x,y,w)$ results from comparing an observation to *CRS* and *VRS* cost functions.

It is possible to link these primal and dual approaches to scale efficiency. Decomposing $CSCE_i(x,y,w)$ into its technical and allocative components, one obtains that $CSCE_i(x,y,w)$ equals $SCE_i(x,y)$ times some ratio of allocative efficiency components:

$$\begin{aligned} CSCE_i(x,y,w) &= \left[\frac{DF_i(x,y|C,S)}{DF_i(x,y|V,S)}\right]\cdot\left[\frac{AE_i(x,y,w|C)}{AE_i(x,y,w|V)}\right] \\ &= SCE_i(x,y)\cdot\left[\frac{AE_i(x,y,w|C)}{AE_i(x,y,w|V)}\right] \end{aligned} \qquad (3)$$

It immediately follows (Färe et al. (1994)):[15] $CSCE_i(x,y,w) = SCE_i(x,y) \Leftrightarrow AE_i(x,y,w \mid C) = AE_i(x,y,w \mid V)$. Since $OE_i(x,y,w \mid C) \leq OE_i(x,y,w \mid V) \leq 1$, the second ratio in (3) can be smaller, equal or larger than unity:

$$[AE_i(x,y,w \mid C)/AE_i(x,y,w \mid V)] \lesseqgtr 1. \quad (4)$$

Furthermore, since both scale efficiency components are smaller or equal to unity (i.e., $CSCE_i(x,y,w) \leq 1$ and $SCE_i(x,y) \leq 1$), one obtains:

$$CSCE_i(x,y,w) \lesseqgtr SCE_i(x,y). \quad (5)$$

The relations between the decompositions in Definition 1 relative to convex and non-convex technologies can be trivially defined as follows (see also Briec et al. (2004: Lemma 3)).

Proposition 1: *Relations between convex and non-convex decomposition components are:*

1) $OTE_i^C(x,y) \leq OTE_i^{NC}(x,y)$;

2) $TE_i^C(x,y) \leq TE_i^{NC}(x,y)$;

3) $C^C(y,w|C) \leq C^{NC}(y,w|C)$;

4) $C^C(y,w|V) \leq C^{NC}(y,w|V)$;

5) $OE_i^C(x,y,w|C) \leq OE_i^{NC}(x,y,w|C)$;

6) $OE_i^C(x,y,w|V) \leq OE_i^{NC}(x,y,w|V)$.

[15] In fact, $SCE_i(x,y)$ in Färe et al. (1994: p. 84-87) is defined on technologies based on limited data, i.e., using information on cost data and the output vector solely. The reader can consult these authors to have more details on the precise conditions to be met.

It is worthwhile stressing that in case of CRS and a single output, relation 3) becomes: $C^{C}\left(y,w\middle|C\right)=C^{NC}\left(y,w\middle|C\right)$ (see Briec et al. (2004): Proposition 4). Hence, under this condition, relation 5) turns into an equality as well: $OE_i^{C}\left(x,y,w\middle|C\right)=OE_i^{NC}\left(x,y,w\middle|C\right)$. Notice that the scale efficiency components cannot be ordered, because they are ratios of other components:

$$SCE_i^{C}(x,y)\gtreqless SCE_i^{NC}(x,y). \qquad (6)$$

Equally obvious, the cost-based scale efficiency components cannot be ordered, because they are again ratios of other components:

$$CSCE_i^{C}(x,y)\gtreqless CSCE_i^{NC}(x,y). \qquad (7)$$

Any eventual differences between both convex and non-convex $OTE_i(x,y)$ and $OE_i(x,y,w\mid C)$ components can be attributed to convexity. Thus, following Briec et al. (2004), one can straightforwardly define a convexity-related technical efficiency ($CRTE_i$ (x,y)) and cost efficiency ($CRCE_i\,(x,y,w)$) component as a ratio between these convex and non-convex components:

Definition 2: *The convex and non-convex efficiency components based upon constant returns to scale technologies respectively cost functions can be related by:*

1) $CRTE_i\,(x,y)=OTE_i^{C}(x,y)/OTE_i^{NC}(x,y)$;

2) $CRCE_i\,(x,y,w)=OE_i^{C}\left(x,y,w\middle|C\right)/OE_i^{NC}\left(x,y,w\middle|C\right)$.

From Definition 3, it follows that $0 < CRTE_i(x,y) \leq 1$ and $0 < CRCE_i(x,y,w) \leq 1$. When $CRTE_i(x,y) = 1$, then the hypothesis that CRS technologies are convex cannot be rejected. The same reasoning applies to the cost functions. These definitions allow to link non-convex and convex decompositions by means of an identity.[16]

3. TECHNOLOGY AND COST FUNCTION SPECIFICATIONS

3.1. Non-Parametric Technology and Cost Function Models

A unified algebraic representation of convex and non-convex technologies under different returns to scale assumptions is possible as follows (see Briec et al (2004)):

$$T^{\Lambda,\Gamma} = \left\{ (x,y): \quad x \geq \sum_{k=1}^{K} x_k \delta\, z_k,\ y \leq \sum_{k=1}^{K} y_k \delta\, z_k,\ z_k \in \Lambda, \delta \in \Gamma \right\},$$

$$\begin{aligned} &\text{where (i) } \Gamma \equiv \Gamma^{CRS} = \{\delta: \quad \delta \geq 0\}; \\ &\text{(ii) } \Gamma \equiv \Gamma^{VRS} = \{\delta: \quad \delta = 1\}; \\ &\text{(iii) } \Gamma \equiv \Gamma^{NIRS} = \{\delta: \quad 0 \leq \delta \leq 1\}; \\ &\text{(iv) } \Gamma \equiv \Gamma^{NDRS} = \{\delta: \quad \delta \geq 1\}; \text{ and} \\ &\text{where} \quad \text{(i) } \Lambda \equiv \Lambda^{C} = \left\{ \sum_{k=1}^{K} z_k = 1 \text{ and } z_k \geq 0 \right\}, \text{ and} \\ &\text{(ii) } \Lambda \equiv \Lambda^{NC} = \left\{ \sum_{k=1}^{K} z_k = 1 \text{ and } z_k \in \{0,1\} \right\}. \end{aligned} \qquad (8)$$

There is one activity vector (z) operating subject to a non-convexity or convexity constraint and a scaling parameter (δ) allowing for a particular scaling of all observations spanning the technology. This scaling parameter is free under CRS, fixed at 1 under VRS, and smaller than or equal to 1 or larger than or equal to 1 under NIRS respectively NDRS.

[16] In particular, $CRTE_i(x,y)$ links non-convex and convex $OTE_i(x,y)$ components by means of the identity:

$$OTE_i^{C}(x,y) = OTE_i^{NC}(x,y).CRTE_i\,(x,y).$$

A similar identity applies to the $OE_i(x,y,w \mid C)$ components.

Computing the radial input efficiency measure relative to convex technologies in (8) requires solving a non-linear programming problem for each evaluated observation. However, Briec and Kerstens (2006) show how this non-linear problem can be transposed into a linear programming problem. Basically, by substituting $w_k = \delta z_k$ in (8), one can rewrite the sum constraint on the activity vector. Realising that the constraints on the scaling factor δ are in fact integrated into the latter sum constraint, the traditional linear program appears (see their Lemma 2.1 for details).

For the non-convex technologies, non-linear mixed integer programs need to be solved. Podinovski (2004b) simplified this computational complexity by suggesting a way to obtain mixed integer programs for all these technologies. Leleu (2006) takes this one step further by formulating a strategy to obtain linear programming problems. Briec and Kerstens (2006) offer a strategy based on implicit enumeration and indicate that the computational complexity of this enumeration is advantageous compared to the recent proposals of Podinovski (2004b) and Leleu (2006).

Turning to the computation of the cost function relative to convex non-parametric technologies, it is well-known that this involves solving a linear program per observation being evaluated (see Färe et al. (1994)). For the cost functions relative to the non-convex technologies, Briec et al. (2004) have developed implicit enumeration algorithms (see Proposition 3).

3.2. Characterising Returns to Scale Information

Several methods have been proposed in the literature to obtain qualitative information regarding returns to scale. However, as argued in Kerstens and Vanden Eeckaut (1999) none of the existing methods is suitable for use with non-convex technologies.

Starting from the goodness-of-fit method earlier proposed by Färe et al. (1983) for convex technologies, Kerstens and Vanden Eeckaut (1999) generalise this method to suit all (including non-convex) technologies.

Definition 3: *Using $DF_i(x,y)$ and conditional on the optimal projection point, technology is locally characterised by:*

a) CRS $\Leftrightarrow DF_i(x,y \mid \text{CRS}) = \max\{ DF_i(x,y \mid \text{C}), DF_i(x,y \mid \text{NIRS}), DF_i(x,y \mid \text{NDRS}) \}$;

b) IRS $\Leftrightarrow DF_i(x,y \mid \text{NDRS}) = \max\{ DF_i(x,y \mid \text{C}), DF_i(x,y \mid \text{NIRS}), DF_i(x,y \mid \text{NDRS}) \}$; or

c) DRS $\Leftrightarrow DF_i(x,y \mid \text{NIRS}) = \max\{ DF_i(x,y \mid \text{C}), DF_i(x,y \mid \text{NIRS}), DF_i(x,y \mid \text{NDRS}) \}$.

Simplifications of this method have been proposed in Soleimani-damaneh et al. (2006) and Soleimani-damaneh and Reshadi (2007). As demonstrated in Briec et al. (2000), the equivalent method for convex technologies can be simplified as follows (see Proposition 3).

Definition 4: *Using $DF_i(x,y)$ and conditional on the optimal projection point, a convex technology is locally characterised by:*

a) IRS $\Leftrightarrow DF_i(x,y|\text{NDRS}) = \max\{ DF_i(x,y|\text{NIRS}), DF_i(x,y|\text{NDRS}) \}$;

b) CRS $\Leftrightarrow DF_i(x,y|\text{NDRS}) = DF_i(x,y|\text{NIRS}) = \max\{ DF_i(x,y|\text{NIRS}), DF_i(x,y|\text{NDRS}) \}$; or

c) DRS $\Leftrightarrow DF_i(x,y|\text{NIRS}) = \max\{ DF_i(x,y|\text{NIRS}), DF_i(x,y|\text{NDRS}) \}$.

Identification of local economies of scale proceeds in very much the same way. A goodness-of-fit method based on the inclusion of different overall efficiency components estimated relative to different return to scale assumptions can be used (see, e.g., Grosskopf (1986) and Sueyoshi (1999) for details). The same reasoning as above applies to infer local scale economies for non-convex and convex technologies.

3.3. Related Literature Applying Non-Convex Models

Meanwhile, FDH is recognised as a standard technology and a variety of empirical applications in different economic contexts are available in the literature. In the public sector studies include, among others, Dervaux et al. (2009) analysing the performance of intensive care units in Pakistan at the individual patient level, Giménez and Prior (2007) analysing Spanish local government efficiency, and Mairesse and Vanden Eeckaut (2002) gauging museum performance in Belgium. Turning to performance studies in the private sector, Alam and Sickles (2000) examine time series of technical efficiency in the US airline industry for convergence, Cullinane et al. (2005) estimate technical efficiency in the world's leading container ports, among others. FDH has also made an inroad in non-traditional production contexts. For instance, Benslimane and Yang (2007) have identified functional requirements in all phases of the procurement process on commercial websites whereby FDH serves to identify the most efficient design. As another example, in a hedonic pricing context FDH has been proposed as an alternative framework to evaluate the performance of heterogeneous products (see Chumpitaz et al. (2010)).

Use of non-convex models including alternative returns to scale assumptions has been more limited. A few studies have reported efficiency levels based on these models: see, e.g., Destefanis (2003) and Destefanis and Storti (2002). Some studies have reported non-convex scale efficiencies: examples include Cesaroni (2011), De Borger and Kerstens (2000) and Mairesse and Vanden Eeckaut (2002). Cesaroni (2011), De Witte and Marques (2011) as well as Mairesse and Vanden Eeckaut (2002) report non-convex returns to scale information for individual units. A few articles have reported and contrasted convex and non-convex cost function results: examples are Briec et al. (2004), Cummins and Zi (1998) and Grifell-Tatjé and Kerstens (2008).

4. DESCRIPTION OF THE SAMPLES

We employ two secondary data sets for our empirical analysis. The first sample is based on 16 Chilean hydro-electric power generation plants observed on a monthly basis for several years (Atkinson and Dorfman (2009)). Limiting ourselves to the observations for the single year 1997, we can safely ignore any technical change and specify an inter-temporal frontier. This results in a total of 192 observations. There is one output quantity (electricity generated) as wells as the price per unit of output. There are also the prices and quantities of three inputs: labour, capital, and water. Except for the input capital, all remaining flow variables are expressed in physical units. Prices are in current Chilean pesos. Basic descriptive statistics for the inputs and the single output are available in Atkinson and Dorfman (2009). Equally so, more details on these data can be found in Atkinson and Dorfman (2009).

As a second sample we draw upon an unbalanced panel of three years of French fruit producers based on annual accounting data collected in a survey (Ivaldi et al. (1996)). Mainly two criteria were adapted to select the farms: (i) the production of apples must be larger than zero, and (ii) the productive acreage of the orchard must be at least five acres. As a technology, three aggregate inputs are combined to produce two outputs. The three inputs are: (i) capital (including land), (ii) labor, and (iii) materials. The two aggregate outputs are (i) the production of apples, and (ii) an aggregate of alternative products. Also input prices are available. Summary statistics and details on the definitions of all variables are available in Appendix 2 in Ivaldi et al. (1996). Notice that the short length of the panel (only three years) justifies the use of an intertemporal frontier that ignores technical change to assess the technical and scale efficiency of these farmers.

5. EMPIRICAL RESULTS

Summary statistics for both convex and non-convex decomposition results in Definition 1 are presented in Table 1. To respect the multiplicative nature of the decomposition geometric averages are used.

< Table 1 about here >

First, the sources of inefficiency differ clearly between convex and non-convex methods. This can be systematically argued as follows. For the hydro-power plants, scale inefficiency is by far the most important source of poor performance under non-convexity. In a convex setting, scale inefficiency is slightly more important than technical inefficiency on the production side, while a marginal reversion of this ranking appears on the cost side. As for the fruit producers, scale inefficiency is the most important source of poor performance under non-convexity, though less pronounced so on the cost side. Under convexity, scale inefficiency is slightly more important than technical inefficiency on the production side, while the reverse occurs on the cost side. Notice that $OE_i(.\,|\,C)$ is extremely low for the hydro-power plants: this may capture extreme variations in the degree of capacity utilization over the year, or it may be due to outliers. Thus, both decompositions yield different conclusions with respect to the major causes of inefficiency, with slight variations between production and cost perspectives.

Second, more observations are efficient under non-convexity, though this difference seems less pronounced under a cost perspective. Clearly, the number of technically efficient units about triples in the non-convex compared to the convex case for the hydro-power plants, and it is even about nine times higher for the fruit producers. Also the number of efficient observations for the scale and overall technical efficiency components increases substantially under non-convexity, except for the cost perspective among the hydro-power plants (see also infra). Thus, a non-convex CRS model tends to be spanned by more or the same number of observations compared to a convex frontier.[17]

Third, the relations between convex and non-convex components as defined in Definition 2 are clearly respected. To bridge the gap between both decompositions, a convexity related component is added. While $CRTE_i(.)$ and $CRCE_i(.)$ are very substantial for the fruit producers, $CRTE_i(.)$ is rather small and $CRCE_i(.)$ even equals unity for the hydro-power plants. The latter result is due to the single output and the CRS assumption and simply illustrates the theoretical result that $C^C\left(y,w\middle|C\right)=C^{NC}\left(y,w\middle|C\right)$ under these conditions (Briec et al. (2004)). This also explains why the number of efficient observations under these conditions is identical (see above). For the fruit producers, convexity plays an important role in explaining inefficiency. In particular, $CRTE_i(.)$ and $CRCE_i(.)$ amount on average to about 38% and 32% respectively. Given an average convex and non-convex $OTE_i(.)$ score of 0.31 and 0.50 respectively, this means that about 19% of this difference can be entirely attributed to convexity. Thus, in a multiple output setting, it cannot be denied that convexity matters both from a production as well as a cost perspective.

Differences between the densities of these efficiency components can be tested with a test statistic proposed by Li (1996) and refined by Fan and Ullah (1999). This Li test statistic has an important characteristic for our purpose: it is valid for both dependent and independent variables. Dependency is distinctive for frontier estimators,

[17] This should normally make the non-convex models less susceptible to outliers.

since efficiency depends on, e.g., sample size. The null hypothesis states the equality of both convex and non-convex efficiency distributions for a given component. Table 2 summarizes the results obtained per data set and per component. Comparing the convex and non-convex results, one observes that for all cases, except the $OTE_i(.)$ and $OE_i(.\mid C)$ scores in the case of the hydro-power plants, the distributions are significantly different. Thus, we can reject the null hypothesis, except for the hydro-power plants on both the production and cost side: while the latter follows from the result in Briec et al. (2004) (see supra), the former may be a bit surprising. Anyway, it confirms the importance of multiple outputs in differentiating convex and non-convex production and cost approaches. Furthermore, the close similarity in the case of a single output and the CRS assumption may contribute to delude people in thinking this similarity is more general than it really is. Finally, comparing production versus cost approaches again distributions of each efficiency component turn out to be significantly different from one another.

To study the effects on the ranking of individual observations, Spearman rank correlations between the components of both decompositions are computed and reported in Table 2. While rankings are very high for both $OTE_i(.)$ and $OE_i(.\mid C)$, they are much lower for both of its components, except for the hydro-power plants evaluated from a cost perspective. While for the hydro-power plants $SCE_i(.)$ and C$SCE_i(.)$ components have a higher degree of similarity in ranking compared to $TE_i(.)$ and $OE_i(.\mid V)$, the situation is mixed for the fruit producers. Finally, the correlation between production and cost perspectives per component is in general lower compared to the correlation between convex versus non-convex setting, except twice for the French fruit producers.

< Table 2 about here >

Next, it is important to verify whether there exist any differences in the determination of returns to scale for individual observations. Table 3 summarises the results per decomposition. First, the majority of hydro-power plant observations are subjected to increasing returns to scale. The same holds true for the fruit producers. While for the hydro-power plants the convex cost approach reveals a non-negligible share of observations subject to decreasing returns to scale, for the fruit producers it is the non-convex approach that consistently indicates such a non-negligible share. Second, there are more observations with constant returns to scale under non-convexity, except for the cost approach applied to the hydro-power plants. This is in line with expectations.

< Table 3 about here >

A natural question to ask is to which extent these overall results hide any differences between convex and non-convex approaches, which is the main focus of this study. Per data set and per production and cost method, we report in Table 4 the percentages of observations for which the returns and economies to scale classification coincides, as well as the ones for which these classifications diverge. First, consensus on the classification varies between 67.71% and 97.92%, leaving a wide to modest margin of conflict. The extreme case of conflict is obviously the switch from increasing returns (economies) to scale to decreasing returns (diseconomies) to scale, or the reverse. This varies from an almost negligible 0.52% to an impressive 22.40% of cases. Second, while production and cost methods yield an about equal amount of consensus for the fruit producers, both approaches differ substantially for the hydro-power plants.

< Table 4 about here >

Another interesting question is to which extent there exists any differences between classifications based upon production and cost methods conditional on the convexity or non-convexity assumption. Per data set and per convexity or non-convexity assumption, we report in Table 5 the percentages of observations for which the returns and economies to scale classification coincides, as well as the ones for which these classifications diverge. First, consensus on the classification varies between 66.15% and 97.40%. Again, this is a wide to modest margin of conflict. The extreme case of conflict is obviously the switch from increasing returns (economies) to scale to decreasing returns (diseconomies) to scale, or the reverse. This varies from 0.00% to an impressive 22.92%, depending on non-convexity or not. Second, while production and cost methods yield a better amount of consensus for the fruit producers under convexity, for the hydro-power plants the highest amount of diagonal element occurs under non-convexity. Thus, while substantial progress has been made since the first study reporting divergences between primal and dual approaches (Burgess, 1975), we can prudently conclude that these differences do not seem to be conditioned by the convexity assumption.

< Table 5 about here >

6. CONCLUSIONS

This contribution is the first to empirically illustrate the differences between technical and scale efficiencies as well as the differences between the characterization of economies of scale and returns to scale based on convex and non-convex technology and cost function estimations. Using data on French fruit producers as well as Chilean hydro-power plants, we empirically observe rather substantial differences regarding the relative importance of technical and scale efficiencies from both a production and a cost perspective. The sample of hydro-power plants also serves to illustrate the theoretical result that convex and non-convex cost functions coincide under CRS and a single output. While for this sample convex and non-convex overall technical efficiency yields slightly different descriptive statistics, the densities turn out to be identical. However, for all other components, the distributions between convex and non-convex technology and cost function are significantly different. Probably more notable, the characterization of both economies of scale and returns to scale for individual observations turns out to be conditioned by convexity in a non-negligible way.

This finding has potentially important consequences for investment decisions, definitions of capacity utilization notions, and other key economic notions. Therefore, it seems important to empirically explore these differences between convex and non-convex technologies and cost functions further in even greater detail (e.g., also focusing on economies of scope, the impact on mergers and acquisitions, the effect on marginal relationships, etc.). In conclusion, even though theoretically the impact of convexity has been known since at least Jacobsen (1970), it seems to be important to further explore the effects of convexity on key economic value relations in practice. Anyway, evidence has been provided that the impact is non-negligible.

Table 1: Non-Convex and Convex Decompositions

Chilian Hydro-power Plants							
	Non-Convex Decomposition				Convex Decomposition		
	$TE_i(.)$[¶]	$SCE_i(.)$	$OTE_i(.)$	$CRTE_i(.)$	$TE_i(.)$	$SCE_i(.)$	$OTE_i(.)$
Average*	0.5139	0.1439	0.0740	0.9366	0.2976	0.2328	0.0693
Stand.Dev.	0.3247	0.2151	0.2148	0.0645	0.3309	0.2648	0.2101
Minimum	0.1293	0.0268	0.0072	0.7438	0.0605	0.0372	0.0070
% Effic. Obs.	30.73%	3.13%	3.13%	27.08%	11.46%	2.08%	2.08%
	$OE_i(. \mid V)$	$CSCE_i(.)$	$OE_i(. \mid C)$	$CRCE_i(.)$	$OE_i(. \mid V)$	$CSCE_i(.)$	$OE_i(. \mid C)$
Average*	0.1633	0.0828	0.0000	1.0000	0.1083	0.1248	0.0000
Stand.Dev.	0.2891	0.1722	0.1080	0.0000	0.2520	0.2174	0.1080
Minimum	0.0144	0.0181	0.0008	1.0000	0.0085	0.0181	0.0008
% Effic. Obs.	3.13%	0.52%	0.52%	100.00%	2.60%	0.52%	0.52%
French Fruit Producers							
	Non-Convex Decomposition				Convex Decomposition		
	$TE_i(.)$	$SCE_i(.)$	$OTE_i(.)$	$CRTE_i(.)$	$TE_i(.)$	$SCE_i(.)$	$OTE_i(.)$
Average*	0.8210	0.6087	0.4997	0.6200	0.5721	0.5416	0.3098
Stand.Dev.	0.1904	0.2379	0.2804	0.1545	0.1933	0.2589	0.2194
Minimum	0.3590	0.0789	0.0486	0.3713	0.1868	0.0728	0.0481
% Effic. Obs.	45.68%	12.84%	12.84%	2.72%	5.43%	2.22%	2.22%
	$OE_i(. \mid V)$	$CSCE_i(.)$	$OE_i(. \mid C)$	$CRCE_i(.)$	$OE_i(. \mid V)$	$CSCE_i(.)$	$OE_i(. \mid C)$
Average*	0.5754	0.5483	0.3155	0.6830	0.3939	0.5470	0.2154
Stand.Dev.	0.2476	0.2049	0.2186	0.1399	0.1898	0.2435	0.1614
Minimum	0.1337	0.0619	0.0393	0.5205	0.1039	0.0567	0.0364
% Effic. Obs.	15.31%	1.98%	1.98%	13.58%	1.73%	0.49%	0.49%

* Geometric Average
[¶] Arguments have been suppressed for all efficiency components.

Table 2: Spearman Rank Correlation Coefficients between Convex and Non-Convex Decomposition Components and between Production and Cost Perspectives

Chilian Hydro-power Plants			
	$TE_i(.)$	$SCE_i(.)$	$OTE_i(.)$
Convex vs. non-convex	0.930*	0.931*	0.994*
Li test	18.18***	6.98***	0.43
	$OE_i(.\mid V)$	$CSCE_i(.)$	$OE_i(.\mid C)$
Convex vs. non-convex	0.975*	0.988*	1.000*
Li test	7.74***	10.07***	0
Production vs. costs	0.843*	0.902*	0.822*
Li test	52.38***	18.18***	50.22***
French Fruit Producers			
	$TE_i(.)$	$SCE_i(.)$	$OTE_i(.)$
Convex vs. non-convex	0.723*	0.786*	0.933*
Li test	104.70***	4.65***	24.98***
	$OE_i(.\mid V)$	$CSCE_i(.)$	$OE_i(.\mid C)$
Convex vs. non-convex	0.807*	0.625*	0.957*
Li test	26.70***	8.54***	12.62***
Production vs. costs	0.725*	0.631*	0.852*
Li test	53.75***	12.48***	24.73***

* Correlation is significantly different from 0 at the 0.01 level (2-tailed).
Li test: critical values at 1% level = 2.33 (***); 5% level = 1.64 (**); 10% level = 1.28 (*).

Table 3: Returns to Scale and Economies of Scale Results

Chilian Hydro-power Plants			
Production	IRS	CRS	DRS
Non-convex	96.35%	3.13%	0.52%
Convex	96.88%	1.56%	1.56%
Cost	IRS	CRS	DRS
Non-convex	98.44%	0.52%	1.04%
Convex	68.23%	9.38%	22.40%
French Fruit Producers			
Production	IRS	CRS	DRS
Non-convex	74.07%	12.84%	13.09%
Convex	90.37%	1.73%	7.90%
Cost	IRS	CRS	DRS
Non-convex	73.83%	1.98%	24.20%
Convex	93.33%	0.25%	6.42%

Table 4: Returns and Economies of Scale: Differences between Convex and Non-Convex Methods

Chilian Hydro-power Plants			
Production	Convex		
Non-Convex	IRS	CRS	DRS
IRS	95.83%		
CRS	1.04%	1.56%	
DRS	0.52%	0.52%	0.52%
Diagonal elements			97.92%
Non-Diagonal elements			2.08%
Cost	Convex		
Non-Convex	IRS	CRS	DRS
IRS	67.19%		
CRS	9.90%	0.00%	
DRS	22.40%	0.00%	0.52%
Diagonal elements			67.71%
Non-Diagonal elements			32.29%
French Fruit Producers			
Production	Convex		
Non-Convex	IRS	CRS	DRS
IRS	74.07%		
CRS	9.38%	1.48%	
DRS	6.91%	2.22%	5.93%
Diagonal elements			81.48%
Non-Diagonal elements			18.52%
Cost	Convex		
Non-Convex	IRS	CRS	DRS
IRS	95.83%		
CRS	1.04%	1.56%	
DRS	0.52%	0.52%	0.52%
Diagonal elements			80.49%
Non-Diagonal elements			19.51%

Table 5: Returns and Economies of Scale: Differences between Production and Cost Models

Chilian Hydro-power Plants			
Non-Convex	Production		
Cost	IRS	CRS	DRS
IRS	96.35%		
CRS	2.08%	0.52%	
DRS	0.00%	0.52%	0.52%
Diagonal elements			97.40%
Non-Diagonal elements			2.60%
Convex	Production		
Cost	IRS	CRS	DRS
IRS	65.63%		
CRS	10.94%	0.00%	
DRS	22.92%	0.00%	0.52%
Diagonal elements			66.15%
Non-Diagonal elements			33.85%
French Fruit Producers			
Non-Convex	Production		
Cost	IRS	CRS	DRS
IRS	68.64%		
CRS	4.44%	1.98%	
DRS	6.17%	6.42%	12.35%
Diagonal elements			82.96%
Non-Diagonal elements			17.04%
Convex	Production		
Cost	IRS	CRS	DRS
IRS	88.15%		
CRS	0.99%	0.25%	
DRS	6.42%	0.49%	3.70%
Diagonal elements			92.10%
Non-Diagonal elements			7.90%

General Conclusions

1. KEY CONCLUSIONS BY CHAPTER

There is a general consensus that frontier efficiency and productivity analysis has offered over the past few decades valuable new measurement tools to monitor and enhance the performance of any organization. Nevertheless, there is always a need to push the knowledge frontier further and to develop new understanding ignored in preceding research.

Also this PhD thesis illustrates the possibility to assess the performance of organizations from an economic and managerial point of view under some new angles. All methodological developments in each of the chapters as well as the empirical applications have been developed with a focus on managerial relevance and potential utility for future research. Therefore, the next three subsections offer specific conclusions for each of these chapters, highlight some interconnections between chapters, and list some lines for future research.

1.1 Managing a Bank Branch Network's Performance

While the short-run industry model has been rather widely used for evaluating regulatory policies of common pool resource industries, in this chapter we have for the first time applied the short-run industry model based on plant capacity measures to compute the productive performance of bank services in a branch network and illustrated how one could use this model as a managerial tool to obtain better efficiency results via a reallocation plan regarding inputs and/or outputs across the network. By focusing on the short-run industry model and it plant capacity measure, we show how the efficiency at the branch level can change in the short run by distinguishing between excess capacity and inefficiency. In a second stage, we show how managers can explore changing the input and output component at the branch level so as to improve the performance of the branch network either by forcing branches to operate at the best practice level of efficiency, or by tolerating some inefficiency. Obviously, performance gains are lower when inefficiencies are tolerated.

An empirical application to the network of a German savings bank in the year 1998 have explored the results of several scenarios using basic efficiency measures relative to both convex and non-convex technologies. The statistical test of Li (1996) has revealed substantial differences between efficiency measures measured relative to convex and non-convex technologies. The latter technologies fit the data better. The detailed empirical results of the short-run industry model at both the industry and branch level reveal a considerable potential for closing down part of the network while maintaining current service levels. This remains true even under the most conservative estimates of efficiency and capacity (i.e., the ones based on a non-convex technology). Additional scenarios related to the impact of adding restrictions on the number of branches, on personnel transfer, and the fixing of alternative aggregate output targets have also been documented.

Clearly, these few scenarios among many do not exhaust the possibilities to adjust this network model to decision-making needs. Obviously, this model -as any theoretical model- has some limits and its applicability would be enhanced if, for instance, geographical information could be included. Meanwhile, alternative models have been explored aiming at essentially the same goals: improving network performance via a reallocation among units. For instance, Vaz et al. (2010) explore the retail store performance distinguishing between two different levels: the section level and the store level. Using a network DEA model the maximum store sales are determined allowing for reallocations of area among the sections within a store. Clearly, the specificities and similarities between different reallocation models remains to be explored.

1.2. Malmquist and Hicks-Moorsteen Productivity Indices: Documenting Infeasibilities

This chapter describes the differences between the Malmquist and Hicks-Moorsteen (or Malmquist TFP) productivity indexes. Both are productivity indexes used to assess the performance of firms over a certain time span, though it cannot be denied that the former index is far more popular.

Usually in the context of total factor productivity (TFP), these two indexes show slightly different TFP results. This difference varies according to the returns to scale assumption on technologies, according to the convexity hypothesis, and also according to the short-run nature or not of the analysis. As earlier research has shown and our empirical analysis confirms the Malmquist productivity index may well be undefined, as its component input or output distance function may be infeasible under certain data configurations. By contrast, the Hicks-Moorsteen productivity index is always well defined and does not lead to any infeasible result at all under minimal assumptions.

Therefore, our empirical analyses illustrates how much infeasibility one can expect to find when using the Malmquist productivity index under some variations in production assumptions. Undoubtedly, more theoretical and empirical studies are needed to check the influence of infeasibilities on these productivity measures. For instance, in future research, it might be meaningful to study the impact of homotheticity on these results.

1.3 Scale Economies and Returns to Scale in Non-Parametric Models

The final chapter has further explored the rather fundamental differences between traditional convex and non-convex production and cost function models in terms of basic static efficiency decompositions While convex models are far more popular, most researchers probably ignore that the traditional convexity hypothesis has both an impact on technical and scale efficiencies estimated based on either a production or a cost model.

The empirical analysis has convincingly documented the differences between technical and scale efficiencies based on both convex and non-convex technology and cost function estimates. It has also exemplified the differences between the characterization of economies of scale and returns to scale for convex as well as non-convex cost functions and technologies for individual observations. Obviously, for inefficient observations such characterization is conditional on a chosen orientation of measurement.

2. LIMITATIONS AND AVENUES FOR FUTURE RESEARCH

As with all empirically-oriented work, there are always areas that could be improved upon. Such shortcomings may also reveal promising avenues for future research. Apart from some of the specific limitations indicated in each of the three chapters or in the three previous subsections, there remain a few general issues worthwhile mentioning at this concluding stage.

A rather common theme throughout this work has been the impact of the traditional convexity hypothesis on empirical results. It is worthwhile pointing out that still lots of domains of empirical application of applied production analysis remain to be explored with regard to this crucial assumption. For instance, the effect of convexity on assessing eventual benefits from mergers of two or several organizations have not yet been documented in the literature. Equally so, if and how traditional measures of capacity utilization are affected by the convexity assumption remains currently totally unknown.

Ultimately, we think such empirical studies need to be supplemented by surveys or discussions with practitioners (e.g., consultants, managers, etc.) to see how they perceive the legitimacy of the convexity hypothesis.

Bibliography

Abbott, M., B. Cohen (2009) Productivity and Efficiency in the Water Industry, *Utilities Policy*, 17(3-4), 233-244.

Alam, I.M.S., R.C. Sickles (1998) The Relationship Between Stock Market Returns and Technical Efficiency Innovations: Evidence from the US Airline Industry, *Journal of Productivity Analysis*, 9(1), 35-51.

Amel, D., C. Barnes, F. Panetta, C. Salleo (2004) Consolidation and Efficiency in the Financial Sector: A Review of the International Evidence, *Journal of Banking & Finance*, 28(10), 2493-2519.

Asmild, M., J.C. Paradi, J.T. Pastor (2009) Centralized Resource Allocation BCC Models, *Omega*, 37(1), 40-49.

Athanassopoulos, A. (1995) Goal Programming & Data Envelopment Analysis (GoDEA) for Target-Based Multi-Level Planning: Allocating Central Grants to the Greek Local Authorities, *European Journal of Operational Research*, 87(3), 535-550.

Athanassopoulos, A., D. Giokas (2000) The Use of Data Envelopment Analysis in Banking Institutions: Evidence from the Commercial Bank of Greece, *Interfaces*, 30(2), 81-95.

Athanassopoulos, A., A. Soteriou, S.A. Zenios (2001) Disentangling Within- and Between Country Efficiency Differences of Bank Branches, in: P.T. Harker, S.A. Zenios (eds.) *Performance of Financial Institutions: Efficiency, Innovation, Regulation*, Cambridge, Cambridge University Press, 336-363.

Atkinson, S.E., C. Cornwell, O. Honerkamp (2003) Measuring and Decomposing Productivity Change: Stochastic Distance Function Estimation Versus Data Envelopment Analysis, *Journal of Business & Economic Statistics*, 21(2), 284-294.

Atkinson, S.E, J.H. Dorfman, (2009) Feasible Estimation of Firm-Specific Allocative Inefficiency Through Bayesian Numerical Methods, *Journal of Applied Econometrics*, 24(4), 675–697.

Balk, B.M. (2003) The Residual: On Monitoring and Benchmarking Firms, Industries and Economies with Respect to Productivity, *Journal of Productivity Analysis*, 20(1), 5-47.

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

Benslimane, Y., Z. Yang (2007) Linking Commercial Website Functions to Perceived Usefulness: A Free Disposal Hull Approach, *Mathematical and Computer Modelling*, 46(9-10), 1191-1202.

Berger, A.N. (2007) International Comparisons of Banking Efficiency, *Financial Markets, Institutions & Instruments*, 16(3), 119-144.

Berger, A.N., R.S. Demsetz, P.E. Strahan (1999) The Consolidation of the Financial Services Industry: Causes, Consequences, and Implications for the Future, *Journal of Banking and Finance*, 23(2-4), 135-194.

Berger, A., D. Humphrey (1997) The Efficiency of Financial Institutions: International Survey and Directions for Future Research, *European Journal of Operational Research*, 98(2), 175-212.

Berger, A.N., J.H. Leusner, J.J. Mingo (1997) The Efficiency of Bank Branches, *Journal of Monetary Economics*, 40(1), 141-162.

Bjørndal, M., K. Jörnsten (2008) Equilibrium Prices Supported by Dual Price Functions in Markets with Non-Convexities, *European Journal of Operational Research*, 190(3), 768–789.

Bjurek, H. (1996) The Malmquist Total Factor Productivity Index, *Scandinavian Journal of Economics*, 98(2), 303-313.

Bjurek, H., F.R. Førsund, L. Hjalmarsson (1998) Malmquist Productivity Indices: An Empirical Investigation, in: R. Färe, S. Grosskopf, R. Russell (eds) *Index Numbers: Essays in Honour of Sten Malmquist*, Boston, Kluwer, 217-239.

Bogetoft, P., L. Otto (2011) *Benchmarking with DEA, SFA, and R*, Springer.

Bouhnik, S., B. Golany, U. Passy, S.T. Hackman, D.A. Vlatsa (2001) Lower Bound Restrictions on Intensities in Data Envelopment Analysis, *Journal of Productivity Analysis*, 16(3), 241-261.

Bragg, S.M. (2002) *Business Ratios and Formulas: A Comprehensive Guide*, New York, Wiley

Briec, W., K. Kerstens (2004) A Luenberger-Hicks-Moorsteen Productivity Indicator: Its Relation to the Hicks-Moorsteen Productivity Index and the Luenberger Productivity Indicator, *Economic Theory*, 23(4), 925-939.

Briec, W., K. Kerstens (2006) Input, Output and Graph Technical Efficiency Measures on Non-Convex FDH Models with Various Scaling Laws: An Integrated Approach Based upon Implicit Enumeration Algorithms, *TOP*, 14(1), 135-166.

Briec, W., K. Kerstens (2009) Infeasibilities and Directional Distance Functions: With Application to the Determinateness of the Luenberger Productivity Indicator, *Journal of Optimization Theory and Applications*, 141(1), 55-73.

Briec, W., K. Kerstens (2011) The Hicks-Moorsteen Productivity Index Satisfies the Determinateness Axiom, *Manchester School*, 79(4), 765–775.

Briec, W., K. Kerstens, H. Leleu, P. Vanden Eeckaut (2000) Returns to Scale on Nonparametric Deterministic Technologies: Simplifying Goodness-of-Fit Methods Using Operations on Technologies, *Journal of Productivity Analysis*, 14(3), 267-274.

Briec, W., K. Kerstens, P. Vanden Eeckaut (2004) Non-convex Technologies and Cost Functions: Definitions, Duality and Nonparametric Tests of Convexity, *Journal of Economics*, 81(2), 155-192.

Burgess, D.F. (1975) Duality Theory and Pitfalls in the Specification of Technologies, *Journal of Econometrics*, 3(2), 105-121.

Burgess, J., P.W. Wilson (1995) Decomposing Hospital Productivity Changes, 1985-1988: A Nonparametric Malmquist Approach, *Journal of Productivity Analysis*, 6(4), 343-363.

Camp, R.C. (1998) Best Practice Benchmarking: The Path to Excellence, *CMA. Hamilton*, 72(6), 10.

Cassels, J.M. (1937) Excess Capacity and Monopolistic Competition, *Quarterly Journal of Economics*, 51(3), 426-443.

Caves, D.W., L.R. Christensen, W.E. Diewert (1982) The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity, *Econometrica*, 50(6), 1393–1414.

Cesaroni, G. (2011) A Complete FDH Efficiency Analysis of a Diffused Production Network: The Case of the Italian Driver and Vehicle Agency, *International Transactions in Operational Research*, 18(2), 205–229.

Chumpitaz, R., K. Kerstens, N. Paparoidamis, M. Staat (2010) Hedonic Price Function Estimation in Economics and Marketing: Revisiting Lancaster's Issue of "Noncombinable" Goods, *Annals of Operations Research*, 173(1), 145-161.

Coelli, T.J., A. Estache, S. Perelman, L. Trujillo (2003) *A Primer on Efficiency Measurement for Utilities and Transport Regulators*, Washington, World Bank Publications.

Coelli, T., E. Grifell-Tatjé, S. Perelman (2002) Capacity Utilisation and Profitability: A Decomposition of Short-Run Profit Efficiency, *International Journal of Production Economics*, 79(3), 261-278.

Coelli, T., D.S.P. Rao, G.E. Battese (2005) *An Introduction to Efficiency and Productivity Analysis*, 2nd Edition, Berlin, Springer.

Copeland, A., and G. Hall (2011) The Response of Prices, Sales, and Output to Temporary Changes in Demand, *Journal of Applied Econometrics*, 26(2), 232–269.

Cullinane, K., D-W. Song, T. Wang (2005) The Application of Mathematical Programming Approaches to Estimating Container Port Production Efficiency, *Journal of Productivity Analysis*, 24(1), 73-92.

Cummins, J.D., M. Rubio-Misas (2006) Deregulation, Consolidation, and Efficiency: Evidence from the Spanish Insurance Industry, *Journal of Money Credit and Banking*, 38(2), 323-356.

Cummins, D., M. Weiss (2000) Analyzing Firm Performance in the Insurance Industry Using Frontier Efficiency and Productivity Methods, in: G. Dionne (ed) *Handbook of Insurance*, Boston, Kluwer, 767-829.

Cummins, D., H. Zi (1998) Comparison of Frontier Efficiency Methods: An Application to the U.S. Life Insurance Industry, *Journal of Productivity Analysis*, 10(2), 131-152.

De Borger, B., K. Kerstens (2000) What Is Known about Municipal Efficiency? The Belgian Case and Beyond, in: J.L.T. Blank (ed.) *Public Provision and Performance: Contributions from Efficiency and Productivity Measurement*, Amsterdam, Elsevier, 299-330.

Debreu, G. (1951) The Coefficient of Resource Utilization, *Econometrica*, 19(3), 273-292.

Deprins, D., D. Simar, and H. Tulkens (1984) Measuring Labor Efficiency in Post Offices, in: M. Marchand, P. Pestieau and H. Tulkens (eds.) *The Performance of Public Enterprises: Concepts and Measurements*, North Holland, Amsterdam.

Dervaux, B., K. Kerstens, H. Leleu (2000) Remedying Excess Capacities in French Surgery Units by Industry Reallocations: The Scope for Short and Long Term Improvements in

Plant Capacity Utilization, in: J.L.T. Blank (ed) *Public Provision and Performance: Contributions from Efficiency and Productivity Measurement*, Amsterdam, Elsevier, 121-146.

Dervaux, B., H. Leleu, E. Minvielle, V. Valdmanis, P. Aegerter, B. Guidet (2009) Performance of French Intensive Care Units: A Directional Distance Function Approach at the Patient Level, *International Journal of Production Economics*, 120(2), 585-594.

Destefanis, S. (2003) The Verdoorn Law: Some Evidence from Non-parametric Frontier Analysis, in: McCombie, J., M. Pugno, B. Soro (eds) *Productivity Growth and Economic Performance: Essays on Verdoorn's Law*, Basingstoke, Palgrave Macmillan, 136-164.

Destefanis, S., G. Storti (2002) Measuring Cross-Country Technological Catch-Up Through Variable-Parameter FDH, *Statistical Methods & Applications*, 11(1), 109-125.

De Witte, K., R.C. Marques (2011) Big and Beautiful? On Non-Parametrically Measuring Scale Economies in Non-Convex Technologies, *Journal of Productivity Analysis*, 35(3), 213-226.

Diewert, W.E. (1992) The Measurement of Productivity, *Bulletin of Economic Research*, 44(3), 163-198.

Diewert, W.E., A.O. Nakamura (1999) Benchmarking and the Measurement of Best Practice Efficiency: An Electricity Generation Application, *Canadian Journal of Economics*, 32(2), 570-588.

Diewert, W.E., A.O. Nakamura (2003) Index Number Concepts, Measures and Decompositions of Productivity Growth, *Journal of Productivity Analysis*, 19(2-3), 127–159.

Elyasiani, E., S. Mehdian, R. Rezvanian (1994) An Empirical Test of Association between Production and Financial Performance: The Case of the Commercial Banking Industry, *Applied Financial Economics*, 4(1), 55-59.

Emrouznejad, A. (2005) Measurement Efficiency and Productivity in SAS/OR, *Computers and Operations Research*, 32(7), 1665-1683.

Epstein, M., J. Henderson (1989) Data Envelopment Analysis for Managerial Control and Diagnosis, *Decision Sciences*, 20(1), 90-119.

Epure, M., K. Kerstens, D. Prior (2011) Technology-Based Total Factor Productivity and Benchmarking: New Proposals and an Application, *Omega*, 39(6), 608–619.

Estache, A., S. Perelman, L. Trujillo (2007) Measuring Quantity-Quality Trade-Offs in Regulation: The Brazilian Freight Railways Case, *Annals of Public and Cooperative Economics*, 78(1), 1-20.

Fan, Y., A. Ullah (1999) On Goodness-of-fit Tests for Weakly Dependent Processes Using Kernel Method, *Journal of Nonparametric Statistics*, 11(1), 337–360.

Färe, R. (1984) The Existence of Plant Capacity, *International Economic Review*, 25(1), 209-213.

Färe, R., S. Grosskopf, E. Kokkelenberg (1989) Measuring Plant Capacity, Utilization and Technical Change: A Nonparametric Approach, *International Economic Review*, 30(3), 655-666.

Färe, R., S. Grosskopf, S-K. Li (1992) Linear Programming Models for Firm and Industry Performance, *Scandinavian Journal of Economics*, 94(4), 599-608.

Färe, R., S. Grosskopf, B. Lindgren, P. Roos (1995) Productivity Developments in Swedish Hospitals: A Malmquist Output Index Approach, in: A. Charnes, W.W. Cooper, A.Y. Lewin, and L. M. Seiford (eds) *Data Envelopment Analysis: Theory, Methodology and Applications*. Kluwer, Boston, 253-272.

Färe, R., S. Grosskopf, C.A.K. Lovell (1983) The Structure of Technical Efficiency, *Scandinavian Journal of Economics*, 85(2), 181-190.

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

Färe, R., S. Grosskopf, D. Margaritis (2008) Efficiency and Productivity: Malmquist and More, in: H. Fried, C.A.K. Lovell, S. Schmidt (eds) *The Measurement of Productive Efficiency and Productivity Change*, New York, Oxford University Press, 522-621.

Farrell, M. (1957) The Measurement of Productive Efficiency, *Journal of the Royal Statistical Society Series A: General*, 120(3), 253-281.

Farrell, M.J. (1959) The Convexity Assumption in the Theory of Competitive Markets, *Journal of Political Economy*, 67(4), 377-391.

Feroz, E.H., S. Kim, R.L. Raab (2003) Financial Statement Analysis: A Data Envelopment Analysis Approach, *Journal of the Operational Research Society*, 54(1), 48-58.

Fisher, I. (1922) *The Making of Index Numbers*, Boston, Houghton-Mifflin.

Førsund, F., L. Hjalmarsson (1974) On the Measurement of Productive Efficiency, *Swedish Journal of Economics*, 76(2), 141-154.

Førsund, F., L. Hjalmarsson (1979) Generalised Farrell Measures of Efficiency: An Application to Milk Processing in Swedish Dairy Plants, *Economic Journal*, 89(354), 294-315.

Førsund, F., L. Hjalmarsson (1983) Technical Progress and Structural Change in the Swedish Cement Industry 1955-1979, *Econometrica*, 51(5):1449–67.

Fried, H.O., C.A.K. Lovell, P. Vanden Eeckaut (1995) Service Productivity in U.S Credit Unions, in: P.T. Harker (ed) *The Service Productivity and Quality Challenge*, Boston, Kluwer, 365-390.

Giménez-García, V.M., J.L. Martínez-Parra, F.P. Buffa (2007) Improving Resource Utilization in Multi-Unit Networked Organizations: The Case of a Spanish Restaurant Chain, *Tourism Management*, 28(1), 262-270**.**

Giménez, V.M., D. Prior (2007) Long- and Short-Term Cost Efficiency Frontier Evaluation: Evidence from Spanish Local Governments, *Fiscal Studies*, 28(1), 121–139.

Goddard, J.A., P. Molyneux, J.O.S. Wilson (2001) *European Banking: Efficiency, Technology and Growth*, New York, Wiley.

Golany, B., E. Tamir (1995) Evaluating Efficiency-Effectiveness-Equality Trade-offs: A Data Envelopment Analysis Approach, *Management Science*, 41(7), 1172-1184.

Grifell-Tatjé, E., K. Kerstens (2008) Incentive Regulation and the Role of Convexity in Benchmarking Electricity Distribution: Economists versus Engineers, *Annals of Public and Cooperative Economics*, 79(2), 227-248.

Grifell-Tatjé, E., C.A.K. Lovell (1995) A Note on the Malmquist Productivity Index, *Economics Letters*, 47(2), 169-175.

Grifell-Tatjé, E., C.A.K. Lovell (1999) A Generalized Malmquist Productivity Index, *TOP*, 7(1), 81-103.

Grosskopf, S. (1986) The Role of the Reference Technology in Measuring Productive Efficiency, *Economic Journal*, 96(382), 499-513.

Harker, P.T., S.A. Zenios (2001) What Drives the Performance of Financial Institutions?, in: P.T. Harker, S.A. Zenios (eds.) *Performance of Financial Institutions: Efficiency, Innovation, Regulation*, Cambridge, Cambridge University Press, 3-31.

Hildenbrand, W. (1981) Short-Run Production Functions based on Microdata, *Econometrica*, 49(5), 1095–1125.

Hirtle, B. (2007) The Impact of Network Size on Bank Branch Performance, *Journal of Banking & Finance*, 31(12), 3782–3805.

Hughes, J.P., L.J. Mester (2008) Efficiency in Banking: Theory, Practice, and Evidence, in: A.N. Berger, P. Molyneux, J. Wilson (eds) *The Oxford Handbook of Banking*, Oxford, Oxford University Press, 463-485.

Ivaldi, M., N. Ladoux, H. Ossard, M. Simioni (1996) Comparing Fourier and Translog Specifications of multiproduct Technology: Evidence from an Incomplete Panel of French Farmers, *Journal of Applied Econometrics*, 11(6), 649-668.

Jacobsen, S.E. (1970) Production Correspondences, *Econometrica*, 38(5), 754–771.

Jamasb, T., P. Nillesen, M. Pollitt (2003) Gaming the Regulator: A Survey, *Electricity Journal*, 16(10), 68-80.

Jamasb, T., M. Pollitt (2003) International Benchmarking and Regulation: An Application to European Electricity Distribution Utilities, *Energy Policy*, 31(15), 1609-1622.

Jehle, G.A., P.J. Reny (2000) *Advanced Microeconomic Theory*, 2nd Ed., Reading, Addison-Wesley.

Johansen, L. (1968) Production Functions and the Concept of Capacity, Namur, Recherches Récentes sur la Fonction de Production (Collection "Economie Mathématique et Econometrie", no 2). Reprinted in: Førsund, F.R. (ed.) (1987) *Collected Works of Leif Johansen, Volume 1*, Amsterdam, North Holland, 359–82.

Johansen, L. (1972) *Production Functions: An Integration of Micro and Macro, Short Run and Long Run Aspects*, Amsterdam, North Holland.

Kantor, J., S. Maital (1999) Measuring Efficiency by Product Group: Integrating DEA with Activity-Based Accounting in a Large Mideast Bank, *Interfaces*, 29(3), 27-36.

Kerstens, K., P. Vanden Eeckaut (1999) Estimating Returns to Scale Using Nonparametric Deterministic Technologies: A New Method Based on Goodness-of-Fit, *European Journal of Operational Research*, 113(1), 206-214.

Kerstens, K., N. Vestergaard, D. Squires (2006) A Short-Run Johansen Industry Model for Common-Pool Resources: Planning a Fishery's Industrial Capacity to Curb Overfishing, *European Review of Agricultural Economics*, 33(3), 361-389.

Klein, L.R. (1960) Some Theoretical Issues in the Measurement of Capacity, *Econometrica*, 28(2), 272-286.

Koopmans, T. (1951) Analysis of Production as an Efficient Combination of Activities, in: T. C. Koopmans (ed) *Activity Analysis of Production and Allocation*, New Haven, Yale University Press, 33-97.

Korhonen, P., M. Syrjänen (2004) Resource Allocation Based on Efficiency Analysis, *Management Science*, 50(8), 1134-1144.

Kumar, S. (2006) Environmentally Sensitive Productivity Growth: A Global Analysis Using Malmquist–Luenberger Index, *Ecological Economics*, 56(2), 280–293.

Leleu, H. (2006) A Linear Programming Framework for Free Disposal Hull Technologies and Cost Functions: Primal and Dual Models, *European Journal of Operational Research*, 168(2), 340-344.

Li, S.K., Y.C. Ng (1995) Measuring the Productive Efficiency of a Group of Firms, *International Advances in Economic Research*, 1(4), 377-390.

Li, Q. (1996) Nonparametric Testing of Closeness between Two Unknown Distribution Functions, *Econometric Reviews*, 15(1), 261–274.

Lozano, S., G. Villa (2004) Centralized Resource Allocation Using Data Envelopment Analysis, *Journal of Productivity Analysis*, 22(1-2), 143-161.

Mairesse, F., P. Vanden Eeckaut (2002) Museum Assessment and FDH Technology: Towards a Global Approach, *Journal of Cultural Economics*, 26(4), 261-286.

Mass-Collell, A., M. Whinston, J. Green (1995) *Microeconomic Theory*, Oxford, Oxford University Press.

McEachern, D., J.C. Paradi (2007) Intra- and Inter-Country Bank Branch Assessment Using DEA, *Journal of Productivity Analysis*, 27(2), 123-136.

Mukherjee, K., S.C. Ray, S.M. Miller (2001) Productivity Growth in Large US Commercial Banks: The Initial Post-Deregulation Experience, *Journal of Banking and Finance*, 25(5), 913-939.

Nash, C., A. Smith (2008) Modelling Performance: Rail, in: D.A. Hensher, K.J. Button (eds) *Handbook of Transport Modelling*, 2nd Ed., Amsterdam, Elsevier, 695-691.

Nelson, R. (1989) On the Measurement of Capacity Utilization, *Journal of Industrial Economics*, 37(3), 273-286.

Nemoto, J., M. Goto (2005) Productivity, Efficiency, Scale Economies and Technical Change: A New Decomposition Analysis of TFP Applied to the Japanese Prefectures, *Journal of the Japanese and International Economies*, 19(4), 617-634.

Nishimizu, M., J. Page (1982) Total Factor Productivity Growth, Technological Progress and Technical Efficiency Change: Dimensions of Productivity Change in Yugoslavia, 1965-78, *Economic Journal*, 92(368), 920-936.

O'Donnell, C.J. (2008) An Aggregate Quantity-Price Framework for Measuring and Decomposing Productivity and Profitability Change, Centre for Efficiency and Productivity Analysis Working Papers WP07/2008, University of Queensland.

O'Donnell, C.J. (2010) Measuring and Decomposing Agricultural Productivity and Profitability Change, *Australian Journal of Agricultural and Resource Economics*, 54(4), 527-560.

Olesen, O., N. Petersen (1996) A Presentation of GAMS for DEA, *Computers and Operations Research*, 23(4), 323-339.

Oral, M., R. Yolalan (1990) An Empirical Study on Measuring Operating Efficiency and Profitability of Bank Branches, *European Journal of Operational Research*, 46(3), 282-294.

Orea, L. (2002) Parametric Decomposition of a Generalized Malmquist Productivity Index, *Journal of Productivity Analysis*, 18(1), 5-22.

Ouellette, P., V. Vierstraete (2004) Technological Change and Efficiency in the Presence of Quasi-Fixed Inputs: A DEA Application to the Hospital Sector, *European Journal of Operational Research*, 154(3), 755-763.

Ouellette, P., V. Vierstraete (2010) Malmquist Indexes with Quasi-Fixed Inputs: An Application to School Districts in Québec, *Annals of Operations Research*, 173(1), 57–76.

Ozcan, Y.A. (2008) *Health Care Benchmarking and Performance Evaluation: An Assessment using Data Envelopment Analysis (DEA)*, Berlin, Springer.

Paradi, J.C., S. Vela, Z. Yang (2004) Assessing Bank and Bank Branch Performance: Modeling Considerations and Approaches, in: W.W. Cooper, L.M. Seiford, J. Zhu (eds) *Handbook on Data Envelopment Analysis*, Kluwer, Boston, 349–400.

Park, J.-B., Y.-W. Jeong, J.-R. Shin, K.Y. Lee (2010). An Improved Particle Swarm Optimization for Nonconvex Economic Dispatch Problems, *IEEE Transactions on Power Systems*, 25(1), 156-166.

Parkan, C. (1987) Measuring the Efficiency of Service Operations: An Application to Bank Branches, *Engineering Cost and Production Economics*, 12(1-4), 237-242.

Pastor, J.T., M. Asmild, C.A.K. Lovell (2011) The Biennial Malmquist Productivity Change Index, *Socio-Economic Planning Sciences*, 45(1), 10-15.

Podinovski, V.V. (2004a) Local and Global Returns to Scale in Performance Measurement, *Journal of the Operational Research Society*, 55(2), 170-178.

Podinovski, V.V. (2004b) On the Linearisation of Reference Technologies for Testing Returns to Scale in FDH Models, *European Journal of Operational Research*, 152(3), 800-802.

Porembski, M., K. Breitenstein, P. Alpar (2005) Visualizing Efficiency and Reference Relations in Data Envelopment Analysis with an Application to the Branches of a German Bank, *Journal of Productivity Analysis*, 23(2), 203-221.

Prior, D. (2003) Long- and Short-Run Non-Parametric Cost Frontier Efficiency: An Application to Spanish Savings Banks, *Journal of Banking & Finance*, 27(4), 655-671.

Ray, S.C. (2004) *Data Envelopment Analysis: Theory and Techniques for Economics and Operations Research*, Cambridge, Cambridge University Press.

Ray, S.C., E. Desli (1997) Productivity Growth, Technical Progress, and Efficiency Change in Industrialized Countries: Comment, *American Economic Review*, 87(5), 1033-1039.

Ray, S.C., K. Mukherjee (1996) Decompositions of the Fisher Ideal Index of Productivity: A Non-Parametric Dual Analysis of US Airlines Data, *Economic Journal*, 106(439), 1659-1678.

Romer, P.M. (1990) Are Nonconvexities Important for Understanding Growth?, *American Economic Review*, 80(2), 97-103.

Scarf, H.E. (1986) Testing for Optimality in the Absence of Convexity, in: W.P. Heller, R.M. Starr, S.A. Starrett (eds) *Social Choice and Public Decision Making: Essays in Honor of Kenneth J. Arrow*, *Volume I*, Cambridge University Press, Cambridge.

Scarf, H.E. (1994) The Allocation of Resources in the Presence of Indivisibilities, *Journal of Economic Perspectives*, 8(4), 111-128.

Seitz, W.D. (1970) The Measurement of Efficiency Relative to a Frontier Production Function, *American Journal of Agricultural Economics*, 52(4), 505-511.

Seitz, W.D. (1971) Productive Efficiency in the Steam-Electric Generating Industry, *Journal of Political Economy*, 79(4), 878-886.

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

Sherman, D., G. Ladino (1995) Measuring Bank Productivity Using Data Envelopment Analysis (DEA), *Interfaces*, 25(2), 60-73.

Silva Portela, M.C.A., E. Thanassoulis (2006) Malmquist Indexes using a Geometric Distance Function (GDF): Application to a Sample of Portuguese Bank Branches, *Journal of Productivity Analysis*, 25(1-2), 25-41.

Simar, L., P.W. Wilson (2008) Statistical Inference in Nonparametric Frontier Models: Recent Developments and Perspectives, in: H. Fried, C.A.K. Lovell, S. Schmidt (eds) *The Measurement of Productive Efficiency and Productivity Change*, New York, Oxford University Press, 421-521.

Soleimani-damaneh, M., G.R. Jahanshahloo, M. Reshadi (2006) On the Estimation of Returns-to-Scale in FDH Models, *European Journal of Operational Research*, 174(2), 1055-1059.

Soleimani-damaneh, M., M. Reshadi (2007) A Polynomial-Time Algorithm to Estimate Returns to Scale in FDH Models, *Computers & Operations Research*, 34(7), 2168-2176.

Sueyoshi, T. (1999) DEA Duality on Returns to Scale (RTS) in Production and Cost Analyses: An Occurrence of Multiple Solutions and Differences between Production-Based and Cost-Based RTS Estimates, *Management Science*, 45(11), 1593-1608.

Thanassoulis, E., M.C.S. Portela, O. Despić (2008) DEA – The Mathematical Programming Approach to Efficiency Analysis, in: H. Fried, C.A.K. Lovell, S. Schmidt (eds) *The Measurement of Productive Efficiency and Productivity Change*, New York, Oxford University Press, 251-420.

Tulkens, H., A. Malnero (1996) Nonparametric Approaches to the Assessment of the Relative Efficiency of Bank Branches, in: Mayes, D. (ed) *Sources of Productivity Growth*, Cambridge, Cambridge University Press, 223-244.

Vaz, C.B., A.S. Camanho, R.C. Guimarães (2010) The Assessment of Retailing Efficiency Using Network Data Envelopment Analysis, *Annals of Operations Research*, 173(1), 5-24.

Voss, C.A., P. Åhlström, K. Blackmon (1997) Benchmarking and Operational Performance: Some Empirical Results, *International Journal of Operations & Production Management*, 17(10), 1046-1058.

Wilson, P.W. (2008) FEAR: A Software Package for Frontier Efficiency Analysis with R, *Socio-Economic Planning Sciences*, 42(4), 247-254.

Worthington, A.C. (2001) An Empirical Survey of Frontier Efficiency Measurement Techniques in Education, *Education Economics*, 9(3), 245-268.

Yörük, B.K., O. Zaim (2005) Productivity Growth in OECD Countries: A Comparison with Malmquist Indices, *Journal of Comparative Economics*, 33(2), 401-420.

Zaim, O. (2004) Measuring Environmental Performance of State Manufacturing through Changes in Pollution Intensities: A DEA Framework, *Ecological Economics*, 48(1), 37–47.

Zaim, O. (2006) A Framework for Incorporating Environmental Indicators to the Measurement of Human Well-Being, in: M. McGillivray (ed.) *Inequality, Poverty and Well-being*, Basingstoke, Palgrave-Macmillan, 194-207.

Zhou, P., B.W. Ang (2008) Decomposition of Aggregate CO_2 Emissions: A Production-Theoretical Approach, *Energy Economics*, 30(3), 1054-1067.

Zofío, J.L. (2007) Malmquist Productivity Index Decompositions: A Unifying Framework, *Applied Economics*, 39(18), 2371–2387.