Nonparametric production and frontier analysis: applications in economics

Marijn Verschelde

Supervisors: Prof. Dr. Glenn Rayp (Ghent University, Supervisor)
Prof. Dr. Koen Schoors (Ghent University, Co-supervisor)

Submitted at Ghent University,
To the Faculty of Economics and Business Administration,
In fulfillment of the requirements for the degree of Doctor in Economics

Doctoral committee: Prof. Dr. Marc De Clercq (Ghent University, Dean)
Prof. Dr. Patrick Van Kenhove (Ghent University, Academic Secretary)
Prof. Dr. Glenn Rayp (Ghent University, Supervisor)
Prof. Dr. Koen Schoors (Ghent University, Co-supervisor)
Prof. Dr. Gerdie Everaert (Ghent University)
Prof. Dr. Meryem Fethi (University of Leicester)
Prof. Dr. Kristiaan Kerstens (CNRS-LEM, IESEG School of Management (Lille))
Prof. Dr. Dirk Van de gaer (Ghent University)

This work was completed as PhD fellow of the Fund
for Scientific Research Flanders (FWO-Vlaanderen)

Submitted at Ghent University,
To the Faculty of Economics and Business Administration,
In fulfillment of the requirements for the degree of Doctor in Economics

Acknowledgments

It is difficult to overstate my gratitude to the many people who helped, inspired and challenged me to complete this academic challenge.

First and foremost I offer my sincerest gratitude to my supervisor, Professor Glenn Rayp. Besides the continuous support he offered me, he was a role model of what an academic researcher should be: precise and honest in every argumentation, serious about what is false and true knowledge, and above all, eager to share his expertise. I wish to thank Glenn in particular for providing me the perfect balance between freedom and guidance. Further, I am truly indebted and thankful for the many insights and fresh new ideas that my co-supervisor Professor Koen Schoors shared with me. Few would disagree that Koen is the personalization of a successful link between academic research and real life policy challenges. I owe sincere and earnest thankfulness to both supervisors, for their help and support in writing this dissertation, but also for teaching me to question even the most widely accepted ideas.

Second, the studies gathered in this dissertation are the result of very inspiring joint research with academics I deeply respect. It is a pleasure to thank my co-authors. Especially, I would like to thank Professor Jean Hindriks for our joint wind ahead journey in combating inefficiencies and inequalities of opportunity in education in Belgium. Further, I would like to express my gratitude to Professor Marijke D'Haese, for inspiring me to deal with maybe the most urgent policy issue in today's world: agricultural development in rural Africa. Additionally, I wish to thank Professor Nicky Rogge and Professor Kristof De Witte, for the many entertaining discus-

sions on conferences and for sharing their eagerness to obtain insight in the operating process of production units of all sorts.

Third, I cannot overstate the thankfulness I feel that I could share both undergraduate and graduate studies with best friends. Thanks Joost, Klaas and Sietse for helping me in this academic quest, for useful comments, but above all, for your friendship. In addition, I was lucky to get to know new friends that enriched my life with humour, sympathy and joy. In particular, lunch time was and is the perfect getaway from academic research. Further, I wish to thank all the colleagues from the department of general economics, for whom I have great regard, not only for insightful discussions and small talk, but also for introducing me to the best pies in Ghent and surroundings. In particular, I want to express my gratitude to Eva and Martine, to always be available for a supportive talk and to guide me through the maze of practicalities that I encountered during the past years.

Finally, I would like to express my heartfelt thanks to my beloved parents, family and friends for their help and wishes for the successful completion of this project. Needless to mention the gratitude I feel to my brother Pieter, as he is always open for a statistical discussion and introduced me to the wonderful world of R. Further, few have the luck as I do to have met at the young age of sixteen the love of their life. Stefanie, my partner in life, I wish to thank you for your support, patience and endless love. I cannot describe the love I feel for you and our baby daughter Lieke. You two complete me. Most importantly, I wish to thank my parents. By pinpointing the importance of education to be able to choose freely in life, you helped me to reach this point. Further, you were always there for me when I needed advice or help. Regrettably, we now know as a family the true meaning of scarcity in time. Therefore, mom, I dedicate this dissertation to you as your love, optimism and energy inspires us all.

Ghent, June 2012

Marijn Verschelde

Table of Contents

List of Figures

List of Tables

1

Introduction[1]

1.1 Overview

Since "*all models are wrong*" (Box, 1976), the art of econometrics consists of choosing and developing models that fit the economic purpose. Few would argue that in the real world, linear economic relations and normal distributions exist, however, these assumptions can lead to useful approximations of true economic relationships (Box, 1976). In fact, economic theory rarely dictates a specific functional form. Instead, it denotes which variables are possibly related and stipulates properties of the relationship (e.g. monotonicity, additivity, concavity) (Yatchew, 1998).

[1] Section 1.4 and Section 1.5 are the result of joint work with Kristof De Witte (Maastricht University, K.U. Leuven).

In this dissertation, I analyze the production of Decision Making Units (DMU's), while limiting 'a priori' assumptions on the production process by the use of nonparametric approaches. I selected four production domains which are characterized by large heterogeneity and/or service production. In these production domains, it is especially hard to assume 'divine insight' in the functional form of production.

1) *Small-scale farming in developing countries.* Although there is a vast literature on estimating agricultural production, there is no consensus on the proper form of the production function (Livanis et al., 2009). Characteristic is the debate on the specification of crop response models. For instance, in contrast to neoclassical theory, the yield response to fertilizers is found to be in direct relation to the quantity of the scarcest factor used ("*the law of the minimum*" of von Liebig, implying non-substitution) (see Paris (1992) and Lanzer and Paris (1981)).[2] A flexible approach is thus warranted to understand agricultural production.

Since the spike of the cereal prices in 2007-08 on world markets, there is increasing awareness in development studies of the role of agricultural development to progress out of poverty and food insecurity. The World Bank and major donors renewed their focus on agricultural development (see e.g. The World Development Report 2008). However, it is not clear which role the vast majority of smallholders play in agricultural development. (Wiggins et al., 2010)

The finding of an inverse relationship between farm productivity and farm size in developing countries, referenced in Chapter 2, is sometimes interpreted as an indication for a large role of small-scale farming in development. However, a significant literature (see e.g. Collier and Dercon (2009)), questions the idea of smallholder farming as the main route for agricultural development in developing economies. Literature pinpoints among others the need of overcoming market imperfections, large-scale investments in e.g. irrigation projects and agricultural technology, promoting commercialization and developing non-agricultural production. Further, the recent successful installment of Chinese-owned superfarms in Sub-Saharan Africa to overcome shortages at home contradict with the claim of more efficient small-scale farms in Africa (Collier

[2] In addition, crop response is non-existing when a large quantity of a fertilizer is used (the "*yield plateau*") (Lanzer and Paris, 1981).

and Dercon, 2009). This dissertation provides additional insight.

In Chapter 2, I use a nonparametric approach to investigate this inverse relationship. A kernel regression is used on data of mixed cropping systems to study the determinants of production including different factors that have been identified in the literature as missing variables in the testing of the inverse relationship such as soil quality, location and household heterogeneity. Household data on farm activities and crop production was gathered among 640 households in 2007 in two Northern provinces of Burundi. The results do not reject the findings of an inverse relationship between farm size and productivity. However, I find that scale elasticity varies substantially, i.e. between 0.2 for the smallest farms and 0.8 for the largest farms. The assumption of a constant scale elasticity over the whole size range is rejected.

This implies that the estimates of the magnitude of the inverse relationship in the existing literature can be biased because of the use of an overly restrictive Cobb-Douglas specification of agricultural production. Additionally, as returns to scale are scale dependent, the findings of an inverse relationship in samples of very small-scale and small-scale farms cannot be extrapolated to large-scale farms. Differently put, I show in my nonparametric analysis that the inverse relationship literature provides little policy information on the benefits and drawbacks from moving from small-scale to large-scale farming. The inverse relationship literature only contains information on the productivity of very small-scale and small-scale farms. In other words, my findings challenge policy advice towards a large role of small-scale farming in agricultural development by supporting the argumentation of Lipton (2010) that big farms cannot be considered as linear blowups of small ones. In addition, I find a micro-level 'productivity-food security' trade-off. Smallholders are more efficient, but also more likely to be (severely) food insecure, which contradicts with the idea that 'small is beautiful' in developing agricultural economies.

2) *The production of cognitive knowledge in secondary education.* The production of cognitive knowledge is a complex, hierarchically structured process. Students follow courses within a school and school type, together with other pupils (the peers) to produce (among others) cognitive knowledge which allows them to be successful in later life. Literature on the economics of education (e.g. Figlio (1999), Baker (2001), Heckman et al. (2008), Henderson et al. (2011))

shows that a traditional specification of educational production functions (such as the Cobb-Douglas model) is overly restrictive. Less restrictive, flexible functional forms that can capture the many interactions and possible non-linearities are needed to understand the hierarchical and cumulative production of cognitive skills.

In Chapter 3, I show the effect of school staff autonomy on educational performance. The distinctive feature with the existing literature is that I employ variation in autonomy within the same country and within the same school type to reduce the omitted variables problem. To fully capture the informational advantage of local actors, I define autonomy as the operational empowerment of the school's direction and teachers. The Flemish secondary school system in Belgium is analyzed as it displays large variation in school staff autonomy. I show in a descriptive approach that school autonomy not always trickles down to the lowest level (i.e., the school's direction and teachers). Combining detailed school level and pupil level data from the PISA 2006 study with a semiparametric hierarchical model, I find a strong positive effect of school staff autonomy in budgeting on educational performance. The result is shown to be robust to problems of reverse causality and simultaneity. Quantile regression estimates show that both low and high-performers benefit from school staff autonomy. The findings support policy oriented towards decentralization of school budget allocation to the school direction and teachers in educational systems with an effective accountability system.

In Chapter 4, I study the relationship between school tracking, social segregation and inequality of opportunity to produce cognitive knowledge. Educational tracking is a very controversial issue in education. The pro-tracking group claims that curriculum and teaching better aimed at children's varied interest and skills will foster learning efficacy. The anti-tracking group claims that tracking systems are inefficient and unfair because they hinder learning and distribute learning inequitably. In this study, I provide a detailed within-country analysis of a specific educational system with a long history of early educational tracking between schools, namely the Flemish secondary school system in Belgium. This is an interesting place to look because it provides a remarkable mix of excellence and inequality. Combining evidence from the PISA 2006 data set at the student and school levels with recent statistical methods, I first relate tracking to social segregation; and second, social segregation to educational opportunity (adequately measured). I

show that tracking, social segregation and inequality of opportunity are closely related. In particular, I show that social segregation, which is considered as socially harmful[3], is institutionalized by the school tracking system.

3) *The evaluation of community oriented local police departments.* Public good provision can among others be characterized by budget-maximization (Niskanen, 1971), by joint production of multiple outputs (Darrough and Heineke, 1979), co-production with the costumers (Whitaker, 1980) and environmental heterogeneity. In result, most academics agree that the production process of public goods cannot be modeled by an a priori imposed functional form (De Witte and Geys, 2011). Nonparametric approaches have shown their value to policy evaluation and efficiency estimation in a large literature on public good provision (e.g. De Borger and Kerstens (1996), Ruggiero (1999), Ruggiero (2000), Balaguer-Coll et al. (2007), De Witte and Geys (2011)).

In Belgium, local police operation is characterized by a combination of autonomy and monitoring. Local police departments operate autonomously, but are subject to the control of the Standing Police Monitoring Committee (Committee P). The present institutional setting provides however little public transparency on the functioning of local police departments.

In Chapter 5, I provide a methodology to rank and evaluate the (perceived) effectiveness of local police departments. I start from the fact that hard data alone are not sufficient to evaluate local police effectiveness in the new age of community policing. Citizens can provide useful feedback regarding strengths and weaknesses of police operations. However, citizen satisfaction indicators typically fail to accurately convey the multidimensional nature of local policing and account for characteristics that are non-controllable for the local police departments. To construct a measure of perceived effectiveness of community oriented police corpses that accounts for both multidi-

[3]Besides the discussed relation between social segregation and inequality of educational opportunity, there are numerous disadvantages of social segregation, reviewed and referenced in Gorard (2009). Among others, social segregation in education can lead to between-school inequalities in information, aspiration, educational resources, dependency on the family background, teacher quality, academic culture, perceptions on a 'fair society', the ability to face diversity. Further, social segregation can imply that students in 'poor' schools are more likely to be delinquent or have a feeling of 'not belonging'.

mensional aspects of local policing and exogenous influences, this study advocates the use of a multivariate conditional, robust order-m version of a non-parametric Data Envelopment Analysis approach with no inputs. I show the potential of the method by constructing and analyzing effectiveness indicators of local police forces in Belgium. In particular, I provide an environment-adjusted 'Benefit-of-the-doubt' estimate of the room for improvement. For half of the sampled local police forces, the room for improvement is estimated to be at least 6 percent. For the least effective police department, this is 20 percent. Further, the advocated approach reveals the 'relative strengths' and 'relative weaknesses' of local police departments. In addition, the findings suggest that perceived police effectiveness is significantly conditioned by the demographic and socioeconomic environment. As community oriented policing is top priority in many developed economies, I believe that the advocated methodology can be easily applied in other countries.

4) *The production of services in bank branches.* In a last chapter, I go into the debate on how to evaluate bank branches. Bank branch evaluation is of particular interest as the structure of both the banking system and individual banks are facing fundamental changes. An adequate reorganization and refocusing of individual banks requires a detailed understanding of bank efficiency at the branch level. How branch-level efficiency is best measured remains unclear. First, the bank branch objectives have evolved. While bank branches were in early literature considered as 'convenience outlets' (Berger et al., 1997), they evolved to 'selling outlets' in the nineties (Athanassopoulos, 1998). In recent years, bank branches are fully acknowledged to play a crucial role in building and maintaining customer loyalty. Second, on the estimation methodology. It is not clear whether we should we impose, possibly locally, distributional assumptions on noise and inefficiency or ignore noise altogether.

To capture the role of 'relationship banking' in bank branch operation, I define a production model in accordance with the new objective of the bank branch, i.e., "*to penetrate its market by selling financial products to new costumers, while tying profitable costumers to the bank and delivering services to existing customers*". Further, I use combined information from a deterministic conditional robust frontier approach and a nonparametric stochastic frontier approach. In the combined approach I control for heterogeneity, non-linearities, environmental variables and

uncertain noise and arrive at a robust understanding of 1) the effectiveness of bank branch resources and 2) the bank branch efficiency levels. I demonstrate the potential of the methodology in a study of market efficiency of 717 bank branches of a large bank in Belgium. Additionally, the findings indicate that returns to scale and inefficiency depend on the environment, and particularly on the client profile.

Using the nonparametric toolkit, I obtain new insight in old and persistent economic issues such as: should we promote small-scale farming in developing countries? Does empowering school staff with operational autonomy benefit education quality? Is school tracking associated with inequalities of opportunity between students to produce cognitive knowledge? Additionally, I show how to '*fairly*' evaluate community oriented local police forces which have multiple tasks and operate in a heterogeneous environment. Finally, I shed some light on how to assess bank branch efficiency given that there is no consensus in the literature on the decomposition of noise and inefficiency. By customizing nonparametric approaches to fit the economic purpose, I make a methodological contribution to the operations research literature. By application, I show the value of nonparametric approaches to discover new insights. As such, I contribute not only to the economic literature, but also to the econometric literature as "*the ultimate test of nonparametric approaches resides in their ability to discover new and unusual relationships*"(Yatchew, 1998).

1.2 The educational production function

While it is clear that farm production, the provision of community oriented police services and bank branch service provision can be investigated by production and frontier analysis, the study of educational attainment by production analysis techniques requires further clarification.
As discussed, the production of cognitive knowledge is a complex, hierarchically structured process. Note that both the pupil as the school are DMU's, the first produces cognitive knowledge and the latter high quality education.
I focus on the production of cognitive knowledge at pupil level by studying an educational production function à la Hanushek (2006) (see (1.1)). The educational outcome at time t of a pupil i

with innate ability α_i in school j is considered to be a function of the cumulating effect of family inputs, peer inputs, school resources, school institutional settings and community effects. The school directly affects production of knowledge of a pupil by the school-specific effect of school resources (school-specific as one school can use its resources more effectively/efficiently than another school). In addition, school institutional settings influence the pupil-level production as appropriate institutions are needed to structure incentives and school culture in line with reaching qualitative education provision. Besides the role of schools and school institutional settings (such as school autonomy (see later)) for efficient production of cognitive knowledge, they also influence educational production indirectly by allowing or preventing inequalities of opportunity. As will be shown, literature argues that a '*fair*' educational system provides a '*level playing field*' for pupils. This implies that in a 'fair' society, educational outcomes may not be related to circumstances beyond the control and responsibility of a pupil. In particular, an institutional setting that institutionalizes or creates incentives towards large effects of family background on education production can be considered as 'unfair'. Evaluation of the production of cognitive knowledge implies thus not only an efficiency and effectiveness evaluation at pupil level, but also a 'fairness' evaluation.

$$O^t_{i,j} = f(F^t_{i,j}, P^t_{i,j}, S^t_j, SI^t_j, C^t_j, \alpha_{i,j}) + v^t_{i,j} \tag{1.1}$$

O^t = Educational outcome at time t

F^t = Family inputs cumulative to time t

P^t = Cumulative peer inputs

S^t = Cumulative school resources

SI^t = Cumulative school institutional settings

C^t = Community effects

α = innate ability

v = noise.

1.3 Parametric and nonparametric analysis

In heterogeneous production environments, use of a linear, additive model can result in large specification biases with erroneous inference as result. Not only can a linear additive model fail to capture all nonlinearities in the true model, it can also miss important interactions. For example, imposition of (log-)linearity can be seen as the imposition that all larger production units are linear blow-ups of smaller ones. When there is considerable heterogeneity in scale, this is often not the case. By imposing additivity between inputs and the environment, the model neglects that the effectiveness of resources can depend on the context they are used in. Especially when there is contextual heterogeneity, additivity is a hard assumption. Adding polynomials (e.g. quadratic terms) and interaction terms is one option to reduce the "bias". However, it can lead to high multicollinearity or low degrees of freedom (Henderson and Kumbhakar, 2006).

One other option is the use of "smoothers" and in specific nonparametric regression techniques. Nonparametric approaches do not impose a priori assumptions on the functional form of the production unit in subject (also called '*divine insight*' (Li and Racine, 2007)). In essence, nonparametric inference comes down to localizing the expected outcome $\mathbf{Y}$ of explanatory variables $\mathbf{X}$ ($E[\mathbf{Y}|\mathbf{X} = \mathbf{x}] \approx E[\mathbf{Y}|\mathbf{X}$ close to $\mathbf{x}]$). The trade-off between bias (you miss important features) and variance (the fit is wiggly and influenced by sampling variation) is settled by selecting the optimal level of localization (using data-driven cross-validation by minimizing e.g. the mean squared error). As such, a localized model allows for non-linearities and interactions where needed by selecting the optimal level of localization.

The flexibility that no a priori knowledge on the functional form is needed comes however at a cost. First, for a given sample, there are fewer neighbour observations in high dimensions (number of inputs, outputs and environmental variables in a production model). Consequently, the reliability of the estimates drops dramatically when the dimension of the model increases. In other words, nonparametric approaches converge slower to the 'true model' when the dimensionality of regressors is high (this is the so called '*curse of dimensionality*'). In the extreme, nonparametric inference in a high-dimensional production environment could lead to a sort of

'*nihilism*' - nothing is significant (Yatchew, 1998). However, the recently introduced localizations of categorical and ordered variables as in Racine and Li (2004) are not sensitive to 'the curse of dimensionality'. As such, the reliability of nonparametric inference only depends on the dimension of continuous variables.

Second, the assumption that the set of observations are realizations of independently, identically distributed random variables is hard for the large majority of applications. Various applications are characterized by complex hierarchical data structures. In education, for example, most of the empirical data have a multilevel structure (pupils are nested within classes, classes within schools, schools within regions and school types, etc.). It is shown that it is necessary to include this highly multilevel data structure into the empirical analysis to obtain unbiased estimates (Raudenbush and Bryk, 2002). To obtain reliable estimates of group-level variables in hierarchical settings, random effects are often introduced to control for unobserved group-level characteristics that are not related to the variables in question. A whole parametric apparatus is developed for this purpose, while nonparametric inference in hierarchically structured datasets is often problematic. However, semiparametric approaches are an alternative to do flexible inference in multilevel (also called 'mixed') settings. By the imposition of additivity, it is possible to include random effects as in the generalized additive mixed model as described in Wood (2006).

Third, a vast literature shows the value of instrumental variables to identify and estimate models that contain endogenous regressors. As an instrumental variable regression goes hand in hand with 'a priori' assumptions on the functional form that are "rarely if ever justified by economic theory"(Horowitz, 2011), a nonparametric approach could be a useful alternative. Recently, a nonparametric instrumental variable regression is proposed by Darolles et al. (2011) and Horowitz (2011). However, routines for applied work (e.g. in the package '*np*' in R of Hayfield and Racine (2008)) are at the moment only in a '*beta*' test phase.

In sum, the benefits of using nonparametric approaches are large 1) in a heterogeneous environment, 2) when the number of (continuous) variables is not too high, 3) when no random effects need to be included and 4) no severe simultaneity is expected.

1.4 Production and frontier analysis

Traditionally, the economic interest is in prices, costs, revenues or profits. However, estimation of a cost function, a revenue function or a profit function requires price information. In many applications, no price information is available or existing (e.g. public good provision). A first approach to estimate a production model without the need of price information is investigating the relation between physical output and physical input quantities. One option is the estimation of the 'average' production function and the productivity variation. In particular, the single-output production function is defined by a $n \times 1$ output scalar $\mathbf{Y}$, a $n \times q$ multivariate regressor $\mathbf{X}$ (inputs), a $n \times r$ vector of environmental variables $\mathbf{Z}$ and an additive error ε:

$$Y_i = g(\mathbf{X}_i, \mathbf{Z}_i) + \varepsilon_i \text{ , with } i = 1, ..., n. \tag{1.2}$$

By definition, all productivity variation around the average production function is a residual. Abramovitz (1956) nicely denotes it as "*a measure of ignorance*". In principle, the residual can be attributed to luck, measurement error, differences in production, technology, differences in the scale of operation and differences in operating efficiency (Fried et al., 2008, p 8).

1.4.1 Stochastic frontier analysis

An alternative is the estimation of input and output distance models which are respectively dual forms of the cost and revenue function that do not require price information (see Shephard (1953) and Shephard (1970)). Productive, technical efficiency is the distance between between the observed and optimal levels of output and input.

In other words, technology is defined by a frontier which consists of benchmark practices and inefficiency is a distance to this frontier. The parametric literature started from a deterministic parametric frontier (see Aigner and Chu (1968)). To overcome problems with the deterministic formulation, Aigner et al. (1977) and Meeusen and van Den Broeck (1977) proposed independently a stochastic frontier which decomposes productive efficiency from random noise by Aigner et al. (1977) and Meeusen and van Den Broeck (1977). Consider for the single-output

case the set of i.i.d. random variables $(\mathbf{X}_i, Y_i, \mathbf{Z}_i)$, with $i = 1, ..., n$, with input $\mathbf{X}_i \in \Re_+^p$, output $Y_i \in \Re_+^1$ and environmental variables $\mathbf{Z}_i \in \Re^r$. Typically, the frontier function $r(\mathbf{X}, \mathbf{Z})$ is introduced as in the parametric model of Aigner et al. (1977):

$$\log Y_i = r(\mathbf{X}_i, \mathbf{Z}_i) - u_i + v_i \text{ , with } i = 1, ..., n. \tag{1.3}$$

The inefficiency term $\mathbf{u} \geq 0$ can be e.g. half normal distributed $(\mathbf{u} \sim |N(0, \sigma_{\mathbf{u}}^2(\mathbf{x}, \mathbf{z})|))$, the error term $\mathbf{v}$ is normally distributed $(\mathbf{v} \sim N(0, \sigma_{\mathbf{v}}^2(\mathbf{x}, \mathbf{z})))$ and $\mathbf{u}$ and $\mathbf{v}$ are independent conditionally on $(\mathbf{X}, \mathbf{Z})$.

Traditionally, parametric assumptions are made on the functional form. Consequently, stochastic frontier analysis requires the imposition of 'a priori' knowledge on 1) the functional form of the frontier and 2) the distribution of noise $\mathbf{v}$ and inefficiency $\mathbf{u}$ which may not be supported by economic theory. However, to loosen the assumptions, a nonparametric stochastic frontier as proposed by Kumbhakar et al. (2007) can be constructed by localizing the maximum-likelihood routine. Still, the nonparametric stochastic frontier requires locally parametric distributional assumptions to decompose noise from inefficiency. As referenced and discussed in Fried et al. (2008, p. 141-153), multiple-output production functions can be estimated by imposing homothetic separability. An alternative is the use of polar coordinates as proposed by Simar (2007).

1.4.2 The non-parametric (deterministic) frontier approach

Deterministic frontier approaches neglect the possibility of random noise. The nonparametric deterministic frontier approach starts from the definition of a production set Ψ, frontier $\mathbf{y}^{\partial}$ and inefficiency λ as a distance to a frontier.[4] Assume that producers use a heterogeneous non-negative input vector $\mathbf{X} \in \mathbb{R}_+^p$ to produce a heterogeneous multivariate output vector $\mathbf{Y} \in \mathbb{R}_+^q$. The production set Ψ of feasible input-output combinations can be defined as:

$$\Psi = \left\{ (\mathbf{X}, \mathbf{Y}) \in \mathbb{R}_+^{p+q} | \text{ } \mathbf{X} \text{ can produce } \mathbf{Y} \right\}. \tag{1.4}$$

[4] Although the outline is limited to the output-oriented case, the extension to input-orientation is straightforward.

In estimating Ψ, two different strands have been developed: 1) the traditional *full* frontier estimators and 2) the *robust* frontier estimators. Each of theses are treated in turn.

First, the traditional 'Data Envelopment Analysis' (DEA; Charnes et al., 1978) literature estimates the production set while including all observed input-output combinations. In other words, it estimates the efficiency of observations relatively to a full frontier. Farrell (1957) and Debreu (1951) were the first to acknowledge that the output-oriented efficiency score (i.e., maximization of outputs $\mathbf{y}$ given the observed inputs $\mathbf{x}$) of an observation $(\mathbf{x},\mathbf{y})$ can be obtained as:

$$\lambda(\mathbf{x},\mathbf{y}) = \sup\{\lambda|(\mathbf{x},\lambda\mathbf{y}) \in \Psi\}. \tag{1.5}$$

A value $\lambda(\mathbf{x},\mathbf{y}) = 1$ indicates full technical efficiency (i.e., there are no observations which are able to produce more outputs for the given input set). A $\lambda(\mathbf{x},\mathbf{y}) > 1$ indicates inefficiency, i.e., it is possible to have a radial increase of $\lambda(\mathbf{x},\mathbf{y})$ in all the outputs in order to reach the efficient frontier. For a given level of input and a given output mix, the efficient level of output is given by:

$$\mathbf{y}^{\partial}(\mathbf{x},\mathbf{y}) = \lambda(\mathbf{x},\mathbf{y})\mathbf{y}. \tag{1.6}$$

Free Disposal[5] Hull (FDH) output-oriented inefficiency estimates $\hat{\lambda}_{FDH}(\mathbf{x},\mathbf{y})$, as introduced in Deprins et al. (1984), are obtained by estimating the output-oriented distance to the free disposal hull of observations. If additionally to FDH a convexity assumption is imposed, one obtains the Data Envelopment Analysis (DEA) inefficiency estimates $\hat{\lambda}_{DEA}(\mathbf{x},\mathbf{y})$. Two main drawbacks of DEA and FDH are the sensitivity to the 'curse of dimensionality' and the extreme sensitivity to outliers. This latter is the result of enveloping all data points. One outlier can make the estimation of the frontier biased and inconsistent.[6]

[5] i.e., if $(\mathbf{x},\mathbf{y}) \in \Psi$, then any $(\mathbf{x}',\mathbf{y}')$ such that $\mathbf{x}' \geq \mathbf{x}$ and $\mathbf{y}' \leq \mathbf{y}$ is also in Ψ. Free disposability thus excludes congestion.

[6] Note however that the problem of sensitivity to outliers is not specific to the non-parametric frontier estimation approaches. Outlying observations can also severely bias the he parametric least squares or stochastic frontier estimates. In the words of Koenker and Bassett (1978), "*the extreme sensitivity of the least squares estimator to modest amounts of outlier contamination makes it a very poor estimator in many non-Gaussian, especially long-tailed,situations.*"

Second, to develop a *robust* frontier estimator, Cazals et al. (2002) formulated the distance model in probabilistic terms. Under the assumption of free disposability, probability theory can be used to interpret the efficiency scores. In particular, efficiency can be viewed as the proportional augmentation of output that unit $(\mathbf{x},\mathbf{y}) \in \Psi$ needs to obtain in order to have zero percent probability to be dominated, given the inputs $\mathbf{x}$. Following Cazals *et al.* (2002), this can be algebraically expressed as:

$$\lambda(\mathbf{x},\mathbf{y}) = \sup\{\lambda | S_{\mathbf{Y}|\mathbf{X}}(\lambda\mathbf{y}|\mathbf{x}) > 0\}, \text{ with } S_{\mathbf{Y}|\mathbf{X}} = \text{Prob}(\mathbf{Y} \geq \mathbf{y}|\mathbf{X} \leq \mathbf{x}). \tag{1.7}$$

By replacing in (1.7) the survival function $S_{\mathbf{Y}|\mathbf{X}}$ by its empirical version $\hat{S}_{\mathbf{Y}|\mathbf{X}}$, Free Disposal Hull (FDH) inefficiency estimates $\hat{\lambda}_{FDH}(\mathbf{x},\mathbf{y})$, as introduced in Deprins et al. (1984), are obtained. Again, Data Envelopment Analysis (DEA) inefficiency estimates $\hat{\lambda}_{DEA}(\mathbf{x},\mathbf{y})$ are obtained by imposing in addition a convexity assumption on the production possibility set.

As the traditional FDH and DEA estimators are sensitive to outlying observations, Cazals et al. (2002) proposed a 'partial frontier' that does no longer envelop all observations. As outlying observations have a high probability to lie 'above' an adequately constructed partial frontier, the sensitivity to outliers is limited. Differently put, by not enveloping all data points, we obtain a robust frontier and robust inefficiency estimates. In particular, Cazals et al. (2002) propose the order-m frontier, which is defined as the expected frontier when considering a random set Ψ_m of only $m < n$ random observations with $\mathbf{X} \leq \mathbf{x}$. As atypical observations are not part of the sub sample Ψ_m in every draw, the impact of such observations on the inefficiency score $\lambda_m(\mathbf{x},\mathbf{y})$ is mitigated. If a decision making unit is expected to perform better than m randomly drawn units, it obtains a super-efficient value of $\lambda_m(\mathbf{x},\mathbf{y}) < 1$, otherwise $\lambda_m(\mathbf{x},\mathbf{y}) \geq 1$. Cazals et al. (2002) made clear that order-m inefficiency $\lambda_m(\mathbf{x},\mathbf{y})$ can be defined as a simple univariate integral function which only depends on the conditional survivor function $S_{\mathbf{Y}|\mathbf{X}}$ (see (1.8)). By replacing $S_{\mathbf{Y}|\mathbf{X}}$ by $\hat{S}_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ in (1.8), we obtain the order-m inefficiency estimate $\hat{\lambda}_m(\mathbf{x},\mathbf{y})$.

$$\lambda_m(\mathbf{x},\mathbf{y}) = \int_0^\infty \left[1 - (1 - S_{\mathbf{Y}|\mathbf{X}}(u\mathbf{y}|\mathbf{x}))^m\right] du. \tag{1.8}$$

An alternative is the robust order-α quantile frontier approach of Aragon et al. (2005). The order-α quantile frontier approach considers efficiency as the proportional augmentation of output that

unit $(\mathbf{x},\mathbf{y}) \in \Psi$ needs to have to obtain a (1-α) percent probability to be dominated, given the inputs. $\alpha \in [0,1]$ and close to 1.

$$\lambda(\mathbf{x},\mathbf{y}) = \sup\{\lambda | S_{\mathbf{Y}|\mathbf{X}}(\lambda\mathbf{y}|\mathbf{x}) > (1-\alpha)\}, \text{ with } S_{\mathbf{Y}|\mathbf{X}} = \text{Prob}(\mathbf{Y} \geq \mathbf{y}|\mathbf{X} \leq \mathbf{x}). \tag{1.9}$$

Besides the robustness to outliers, which is shown theoretically and numerically in Daouia and Ruiz-Gazen (2006) and Daouia and Gijbels (2011), these partial frontier approaches are proven to converge by a $\sqrt{n}$-rate to the true partial frontier. In result, we obtain reliable estimates that are less vulnerable for the '*curse of dimensionality*'.

1.5 Capturing the environment: two-stage versus conditional deterministic frontier approach

As most production units not operate in vacuum, it is necessary to control for environmental characteristics outside the control of the decision maker to obtain 'fair' efficiency evaluation. While the environment can be directly included in the specification of a parametric stochastic frontier as in (1.3)[7], conditioning for the environment in a deterministic nonparametric framework is less straightforward. There are multiple approaches for this sake[8]. The far-most popular approach is also the most controversial one: a two-stage approach. The two-stage approach estimates in a first phase non-parametrically the efficiency scores (most commonly by FDH or DEA). In a second phase, it explains the obtained estimates by a parametric regression. Simar and Wilson (2007), Simar and Wilson (2011), "*2-stage DEA: Caveat Emptor*" and Johnson and Kuosmanen (2012) show rigorously that the second-stage inference is invalid in the thousands of studies that use a two-stage approach. Simar and Wilson (2011) advice the use of a conditional efficiency framework to smoothly condition the efficiency estimates on the environment.

[7] See e.g. Kumbhakar and Lovell (2000) for an overview.

[8] For a review see e.g. Daraio and Simar (2007, p. 96-100).

1.5.1 Two-stage approach

Although the two-stage approach has been applied frequently, Simar and Wilson (2007) indicate rigorously that it can be a rather tricky procedure.

First, as presented in Simar and Wilson (2007) and Daraio et al. (2010), a two-stage approach is only justified when a separability condition in the Data Generating Process (DGP) is introduced. To make clear what the separability assumption implies, I follow Daraio et al. (2010) in specifying the DGP in 3 assumptions. For this, let $\mathbf{X} \in \mathbb{R}_+^p$ define a vector of p input quantities, $\mathbf{Y} \in \mathbb{R}_+^q$ denote a vector of q output quantities and $\mathbf{Z} \in \mathbb{R}^r$ denote a vector of r environmental variables. In Assumption 1.5.1, I define how the data is generated. In 1.5.2, I impose the separability assumption. In Assumption 1.5.3, I define how environmental variables $\mathbf{z}$ influence inefficiency.

Assumption 1.5.1 *The sample of n observations* $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)$ *in* $\mathbb{S}_n = \{(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i)\}_{i=1}^n$ *are identically, independently distributed (iid) random variables* $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ *with probability density function* $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$*, which has support over the compact attainable set* $\mathcal{P} \subset \mathbb{R}_+^{p+q} \times \mathbb{R}^r$ *with conditional production sets* $\mathcal{P}(\mathbf{z})$ *defined by*

$$\mathcal{P}(\mathbf{z}) = \{(\mathbf{X}, \mathbf{Y}) | \mathbf{Z} = \mathbf{z}, \mathbf{X} \textit{ can produce } \mathbf{Y}\}. \tag{1.10}$$

Let

$$\Psi = \bigcup_{\mathbf{z} \in \mathcal{Z}} \mathcal{P}(\mathbf{z}). \tag{1.11}$$

Assumption 1.5.2 $\mathcal{P}(\mathbf{z}) = \Psi \; \forall \mathbf{z} \in \mathcal{Z}$.

Assumption 1.5.3 $\mathbf{Z}_i$ *influences* λ_i *through the following mechanism:*

$$\lambda_i = m(\mathbf{Z}_i) + \varepsilon_i \geq 1, \tag{1.12}$$

where m is a smooth, continuous function and ε_i *is a continuous iid random variable, independent of* $\mathbf{Z}_i$.

The separability assumption, defined in Assumption 1.5.2, implies that environmental variables **z** do not affect the support of **y**. In other words, environmental variables **z** have no impact on the position of the frontier. Only inefficiency is affected by environmental variables **z**. In assumption 1.5.3, I specify that the effect of unbounded environmental variables **z** on inefficiency (λ) is truncated.[9]

Second, Simar and Wilson (2007) show that inference is invalid in many papers that use a typical two-stage procedure with traditional efficiency estimates - such as DEA and FDH - in the first stage. $\hat{\lambda}$ is an estimate of relative inefficiency and is is by construction not i.i.d. As inefficiency estimate $\hat{\lambda}$ and not the true inefficiency λ is the dependent variable in the second stage, there is by definition dependency in the model, with biased inference as result. To overcome these problems, Simar and Wilson (2007) propose to use a double bootstrap procedure to construct left-truncated bias-corrected efficiency estimates and make inference valid under the separability assumption. An alternative is a conditional efficiency approach.

1.5.2 Conditional efficiency approach

The conditional efficiency approach - as introduced by Daraio and Simar (2005) and Daraio and Simar (2007b) - uses the probabilistic formulation of efficiency estimations - as introduced by Cazals et al. (2002) - to introduce environmental variables $\mathbf{Z}$ directly in the production process. In contrast to the more traditional two-stage approach, by using a probabilistic formulation, the conditional efficiency approach does not impose a separability assumption between the input $\times$ output space and the space of $\mathbf{Z}$ values as defined in Assumption 1.5.2. In other words, $\mathbf{Z}$ can influence the attainable set Ψ or the position of the frontier of the attainable set. The conditional survival function is expressed as:

$$S(\mathbf{x},\mathbf{y}|\mathbf{z}) = \text{Prob}(\mathbf{Y} \geq \mathbf{y}|\mathbf{X} \leq \mathbf{x}, \mathbf{Z} = \mathbf{z}), \tag{1.13}$$

[9] An alternative to explain efficiency in a second stage approach is a logit regression.

such that the conditional efficiency is obtained as:

$$\lambda(\mathbf{x},\mathbf{y}|\mathbf{z}) = \sup(\lambda|S_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}(\lambda\mathbf{y}|\mathbf{x},\mathbf{z}) > 0). \tag{1.14}$$

An estimator of $\lambda(\mathbf{x},\mathbf{y}|\mathbf{z})$ (i.e., $\hat{\lambda}(\mathbf{x},\mathbf{y}|\mathbf{z})$) can be constructed by smoothing $\mathbf{Z}$ by the use of a kernel estimator. Daraio and Simar (2005, 2007b) presented a framework to visualize the effects of the exogenous variables $\mathbf{Z}$. In particular, they suggested that by regressing non-parametrically $\hat{\lambda}(\mathbf{x},\mathbf{y}|\mathbf{z})$ / $\hat{\lambda}(\mathbf{x},\mathbf{y})$ on $\mathbf{Z}$, the direction of influence can be estimated. Conditional versions of full frontier approaches or partial frontier approaches can be used in a conditional efficiency approach. Therefore, the conditional approach is robust for outliers in $\mathbf{X}$ and $\mathbf{Y}$ when a partial frontier approach is used and smoothly conditions efficiency on the environment.

As discussed, nonparametric approaches have as main advantage the relaxation of parametric assumptions on the functional form of production. A main drawback is the estimation problems when the dimensionality of production is high, compared to the sample size, or hierarchically structured. In addition, in comparison to stochastic frontier approaches, nonparametric deterministic frontier approaches are sensitive to outlying observations and it is less straightforward to capture the environment.

References

Abramovitz, M., 1956. Resource and output trends in the US since 1870. American Economic Review 46 (2), 5–23.

Aigner, D., Chu, S., 1968. On estimating the industry production function. American Economic Review 58 (4), 826–839.

Aigner, D., Lovell, C., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. Journal of Econometrics 6, 21–37.

Aragon, Y., Daouia, A., Thomas-Agnan, C., 2005. Nonparametric frontier estimation: A conditional quantile-based approach. Econometric Theory 21 (2), 358–389.

Athanassopoulos, A., 1998. Nonparametric frontier models for assessing market and cost efficiency of large-scale bank branch networks. Journal of Money, Credit, and Banking 30 (2), 173–192.

Baker, B. D., 2001. Can flexible non-linear modeling tell us anything new about educational productivity? Economics of Education Review 20 (1), 81–92.

Balaguer-Coll, M. T., Prior, D., Tortosa-Ausina, E., 2007. On the determinants of local government performance: A two-stage nonparametric approach. European Economic Review 51 (2), 425–451.

Bank, W., 2007. World Development Report 2008: Agriculture for Development. Washington, DC: The World Bank.

Berger, A., Leusner, J., Mingo, J., 1997. The efficiency of bank branches. Journal of Monetary Economics 40, 141–162.

Box, G., 1976. Science and statistics. Journal of the American Statistical Association 71 (356), 791–799.

Cazals, C., Florens, J. P., Simar, L., 2002. Nonparametric frontier estimation: A robust approach. Journal of Econometrics 106 (1), 1–25.

Charnes, A., Cooper, W. W., Rhodes, E., 1978. Measuring efficiency of Decision-Making Units. European Journal of Operational Research 2 (6), 429–444.

Collier, P., Dercon, S., 2009. African agriculture in 50 years: Smallholders in a rapidly changing world? In: FAO, UN Economic and Social Development Department.

Daouia, A., Gijbels, I., 2011. Robustness and inference in nonparametric partial frontier modeling. Journal of Econometrics 161 (2), 147–165.

Daouia, A., Ruiz-Gazen, A., 2006. Robust nonparametric frontier estimators: Qualitative robustness and influence function. Statistica Sinica 16 (4), 1233–1253.

Daraio, C., Simar, L., 2005. Introducing environmental variables in nonparametric frontier models: A probabilistic approach. Journal of Productivity Analysis 24 (1), 93–121.

Daraio, C., Simar, L., 2007a. Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications. Studies in productivity and efficiency. Springer Science and Business Media.

Daraio, C., Simar, L., 2007b. Conditional nonparametric frontier models for convex and nonconvex technologies: A unifying approach. Journal of Productivity Analysis 28 (1-2), 13–32.

Daraio, C., Simar, L., Wilson, P., 2010. Testing whether two-stage estimation is meaningful in non-parametric models of production. ISBA Discussion Paper (1031).

Darolles, S., Fan, Y., Florens, J. P., Renault, E., 2011. Nonparametric instrumental regression. Econometrica 79 (5), 1541–1565.

Darrough, M., Heineke, J. M., 1979. Law-enforcement agencies as multi-product firms - Econometric investigation of production costs. Public Finance-Finances Publiques 34 (2), 176–195.

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

De Witte, K., Geys, B., 2011. Evaluating efficient public good provision: Theory and evidence from a generalised conditional efficiency model for public libraries. Journal of Urban Economics 69 (3), 319–327.

Debreu, G., 1951. The coefficient of resource utilization. Econometrica 19, 273–292.

Deprins, D., Simar, L., Tulkens, H., 1984. Measuring labor-efficiency in post offices. In: Marchand, M., Pestieau, P., Tulkens, H. (Eds.), The performance of public enterprises - concepts and measurement. Amsterdam, North-Holland, pp. 243–267.

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

Figlio, D. N., 1999. Functional form and the estimated effects of school resources. Economics of Education Review 18 (2), 241–252.

Gorard, S., 2009. Does the index of segregation matter? The composition of secondary schools in England since 1996. British Educational Research Journal 35 (4), 639–652.

Hanushek, E., 2006. School resources. In: Hanushek, E., Welch, F. (Eds.), Handbook of the economics of education. pp. 865–908.

Hayfield, T., Racine, J., S., 2008. np: Nonparametric kernel smoothing methods for mixed datatypes. R package version 0.14-3.

Heckman, J., Lochner, L., Todd, P., 2008. Earnings functions and rates of return. NBER Working Paper Series (13780).

Henderson, D. J., Kumbhakar, S. C., 2006. Public and private capital productivity puzzle: A nonparametric approach. Southern Economic Journal 73 (1), 219–232.

Henderson, D. J., Polachek, S. W., Wang, L., 2011. Heterogeneity in schooling rates of return. Economics of Education Review 30 (6), 1202–1214.

Horowitz, J. L., 2011. Applied nonparametric instrumental variables estimation. Econometrica 79 (2), 347–394.

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

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

Kumbhakar, S., Lovell, C., 2000. Stochastic frontier analysis. Cambridge University Press.

Kumbhakar, S. C., Park, B. U., Simar, L., Tsionas, E. G., 2007. Nonparametric stochastic frontiers: A local maximum likelihood approach. Journal of Econometrics 137 (1), 1–27.

Lanzer, E., Paris, Q., 1981. A new analytical framework for the fertilization problem. American Journal of Agricultural Economics 63 (1), 93–103.

Li, Q., Racine, J., 2007. Nonparametric Econometrics: Theory and practice. Princeton University Press.

Lipton, M., 2010. From policy aims and small-farm characteristics to farm science needs. World development 10, 1399–1412.

Livanis, G., Salois, M., Moss, C., 2009. A nonparametric kernel representation of the agricultural production function: Implications for economic measures of technology, Presented at the 83rd Annual Conference of the Agricultural Economics Society, Dublin 30 March - 1 April, 2009.

Meeusen, W., van Den Broeck, J., 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. International Economic Review, 435–444.

Niskanen, W., 1971. Bureaucracy and Representative Government. Chicago: Aldine-Atherton.

Paris, Q., 1992. The von Liebig hypothesis. American Journal of Agricultural Economics 74 (4).

Racine, J., Li, Q., 2004. Nonparametric estimation of regression functions with both categorical and continuous data. Journal of Econometrics 119 (1), 99–130.

Raudenbush, S. W., Bryk, A. S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods. Thousand Oaks, CA: Sage.

Ruggiero, J., 1999. Nonparametric analysis of educational costs. European Journal of Operational Research 119 (3), 605–612.

Ruggiero, J., 2000. Nonparametric estimation of returns to scale in the public sector with an application to the provision of educational services. Journal of the Operational Research Society 51 (8), 906–912.

Shephard, R., 1953. Cost and Production Functions. Princeton: Princeton University Press.

Shephard, R., 1970. Theory of Cost and Production functions. Princeton: Princeton University Press.

Simar, L., 2007. How to improve the performances of DEA/FDH estimators in the presence of noise? Journal of Productivity Analysis 28 (3), 183–201.

Simar, L., Wilson, P., 2008. Statistical inference in nonparametric frontier models: recent developments and perspectives. In: Fried, H., Lovell, C., Schmidt, S. (Eds.), The measurement of productive efficiency. Oxford University Press.

Simar, L., Wilson, P. W., 2007. Estimation and inference in two-stage, semi-parametric models of production processes. Journal of Econometrics 136 (1), 31–64.

Simar, L., Wilson, P. W., 2011. Two-stage DEA: Caveat emptor. Journal of Productivity Analysis 36 (2), 205–218.

Whitaker, G. P., 1980. Coproduction - Citizen participation in service delivery. Public Administration Review 40 (3), 240–246.

Wiggins, S., Kisten, J., Llambi, L., 2010. The future of small farms. World Development 38, 1341–1348.

Wood, S., 2006. Generalized Additive Models: an introduction with R. Texts in Statistical Science. Chapman and Hall/CRC.

Yatchew, A., 1998. Nonparametric regression techniques in economics. Journal of Economic Literature 36, 669–721.

2

Challenging small-scale farming, a non-parametric analysis of the (inverse) relationship between farm productivity and farm size in Burundi[1]

2.1 Introduction

The (possibly inverse) relationship between farm size and land productivity has been heavily debated in literature for decades (see e.g. Wiggins et al. (2010) and Eastwood et al. (2010)). Given that it contradicts economic theory, which implies that marginal factor productivity should be equal across farms or between the plots of a single household, all along attempts are made to

[1]This chapter is the result of joint work with Marijke D'Haese, Glenn Rayp and Ellen Vandamme. Another version of this chapter appeared as Ghent University Working Paper WP 11/745.

explain the occurrence of this inverse relationship. Several obvious and less obvious reasons and explanations for the inverse relationship between farm size and farm productivity (IR) have been put forward and tested, but none have yet been able to provide an explanation for the IR.

A first obvious reason is the presence of imperfect factor markets (Feder, 1985). This includes failures in different types of production factor markets: land market (Platteau, 1996; Heltberg, 1998), credit market (Assunção and Ghatak, 2003), insurance market (Dercon and Krishnan, 1996) and labour market (Feder, 1985; Barrett, 1996; Assunção and Braido, 2007). Malfunctioning or a complete absence of these markets will lead to suboptimal resource allocation on farm level implying inefficiencies. An important cause of the presence of imperfect labour markets in developing countries is claimed to be labour supervision cost (Feder, 1985; Lipton, 2010). As hired labour is assumed to be less motivated and effective, it takes more productive family labour to supervise hired labour which decreases overall labour productivity at farm level. This would explain why labour and farm productivity are lower on large farms, which require more hired labour. Assunção and Braido (2007) and Barrett et al. (2010) argue that the imperfect market hypotheses imply the presence of unobservable variation between households that leads to differences in the input intensity levels which are correlated with farm area. Therefore, they add a set of household specific characteristics such as household size, dependency ratio, and gender of the household head in testing the inverse relation between farm size and productivity. However, none of the studies cited up to now has proven household characteristics to solely explain the IR.

A second important explanation questions whether the IR between farm size and productivity emerges (or not) due to omitted variables. Soil quality is mentioned as an important but often neglected explanatory variable. Differences in soil quality lead to differences in soil productivity which clearly affect output (Sen, 1975), with small farmers being more productive because of having plots of better quality. In addition, farming practices and production methods might vary according to farm size, leading to differences in yields and productivity (Byiringiro and Reardon, 1996; Schultz, 1964; Assunção and Braido, 2007; Lipton, 2010). All revised studies on this issue show a decrease in the severity of the IR when controlling for soil quality (Benjamin,

1995; Lamb, 2003; Assunção and Braido, 2007; Barrett et al., 2010), but none has found that the IR disappears when controlling for soil quality. Differentiation in farm management skills as an explanatory variable of farm productivity was tested using panel data in which was allowed for household-specific fixed effects (Lipton, 2010). Though Lipton (2010) argues that differentiation in management was not yet thoroughly tested in empirical research, the existing evidence doesn't point to managerial skills as the determinant which explains the IR.

A third explanation of the IR is related to methodological issues. The debate in literature points to the struggles with methodological problems in proving the IR. As one of the unsettled issues, Lipton (2010) mentions that big farms cannot be considered as linear blowups of small ones. Incentives to use inputs vary with the production scale, i.e. bigger farms use a different technology than small farms. Most empirical studies on the IR are based on cross sectional data used for econometric models that fail to capture for non-linearities and that impose a common specification (parameters) for the whole sample they analyze. Moreover, the scale ranges that are allowed in the models may be too small to measure scale effects (Collier and Dercon, 2009).

This paper addresses in particular the latter issue. We analyze factors influencing farm production including scale using a non-parametric estimation of the production function estimated for a unique dataset of farmers in the North of Burundi. Barrett (1996), Assunção and Braido (2007) and Barrett et al. (2010) already suggested the use of a nonparametric regression to show the occurrence of an inverse relationship. However, the nonparametric part of the analysis was limited to an illustrative bivariate kernel or spline regression of yield on (cultivated) land size. A parametric regression was used to investigate possible explanations for the inverse relationship. In this study, we analyze the inverse relationship without imposing any parametric assumption on the specification of the production function. In particular, we use the recently introduced multivariate Racine and Li (2004) generalized kernel regression which allows for both categorical and continuous data. As such, we allow for non-linearity and interactions where needed between both continuous and categorical variables. Specifically, by using a nonparametric approach we are able to track heterogeneity in productivity effects of increased access to production factors.

Our rich dataset allows controlling for several of the missing variables mentioned above. We account for mixed output by valuing food and cash crops produced at opportunity cost and market value. The relationship between inputs and farm output is not linear, which parametric models fail to capture. We find that scale elasticity of the largest farms is fourfold that of a smaller farm. However, for the sample we analyze, we fail to reject the inverse relationship. In addition, we confirm for Burundi the importance of missing variables to which is referred in the literature such as soil quality and field fragmentation. In the next section we describe the data and methods used in the analysis. The third section presents our estimation results. Conclusions are drawn in the fourth section.

2.2 Data

Household data on farm activities was gathered in 2007 in two densely populated provinces in the North of Burundi, Ngozi and Muyinga. The provinces were chosen because they are among the most populated of the country. Both provinces cover an area of 2300 km^2 and 1.4 million inhabitants; this is 13% of the total surface of Burundi and 19% of the population. Both provinces are densely populated with 475 inhabitant per km^2 in Ngozi and 322 inhabitants km^2 in Muyinga. Economic activity outside agriculture is very limited in both provinces, except for the city of Ngozi which is the third largest city of Burundi. Burundi has the sad record of being one of the poorest countries in the world. With a GNI of 390$ (PPP) per capita it is ranked at the bottom of the group of low-income countries (World Bank, 2011). In the Human Development Index ranking of 169 countries, it is at the 166th place (UNDP, 2010). The country seems to have much against it when trying to succeed in promoting economic growth; its size is rather small, it is landlocked, with limited natural resources and it is prone to ethnic conflict. The economy depends largely on agriculture; more than one third of the total GDP is derived from agricultural production and more than 90% of employment is allocated to the agricultural sector. Agriculture also plays a vital role in the trade balance as more than 90% of foreign exchange earnings is derived from the export of coffee although the contribution of this export to the country's GDP is rather small (CIA, 2010).

In total 640 farm households were questioned; 360 in the Nogzi Province and 280 in Muyinga Province. All 16 municipalities of the two provinces were covered (nine in Ngozi Province and seven in Muyinga), per municipality ten villages where selected based on geographical distribution and in every village four households were randomly selected. The interviews were held in Kirundi in collaboration with a team of the University of Burundi. Because of missing data, 20 farms had to be excluded from the data analysis.

For each household, two questionnaires were used; a first questionnaire collected information on household and farm characteristics, including food security issues. A second questionnaire was used to gather information on each plot the farmer owned. The result is a very rich dataset with detailed and reliable information on farm scale (production level, size, labour input, farm inputs), the farming system (crop choices and cash crops) as well as on the farmer's evaluation of the soil quality, and steepness of the different plots. The latter is particularly important given the area is particularly hilly. In order to avoid measurement error in farm size, positions were measured using GPS (GPSmap 60CSx), which allowed for a measurement of the plot size with a precision of 5 meters.

2.2.1 Variables included in the model

Burundi is, as most African agricultural economies, characterized by severe market imperfections. Only 10 % of the sampled farmers denote they have access to credit, indicating severe credit market imperfections. As land market imperfections prevent the selling and buying of land, a large proportion of the land surface is obtained by inheritance. In our sample, on average, 67% of the farm area used for cultivating food and cash crops is inherited. Further, one third of the sampled smallholders obtained all their land by inheritance.[2] In consequence of the low level of marketing in rural Africa, there can be a large price wedge between the value of food crops on local markets and their opportunity costs (Low, 1986). Therefore, where possible, we

[2]Figures not included in Table 2.1, but available upon request.

value food crops by opportunity cost.[3] Opportunity cost is that of avoided imports, plus transport and handling costs to the villages surveyed. We treat transport costs and handling costs as a village-specific lump sum cost for which we control by including community fixed effects into the production model.[4] In specific, we measure food crops at opportunity cost, using CEPII[5] import unit values. Food crops for which no import unit values are available (as import is rare or non-existing) and cash crops are valued at market value.[6] The output of these crops is measured by the sum of the market value of all crops produced irrespective of whether these are sold or consumed by the household. Farm production for each of these crops is multiplied by the average market price of the respective crops. The level of marketing by the farmers is so low that no individual farm-gate prices could be captured.
Furthermore, the diversity of the mixed cropping produce made it impossible to use other quantity measures. E.g., the alternative of caloric content could not be used because it would exclude the possibility to account for the value of coffee production.

Descriptive statistics for all variables included in the model are given in Table 2.1. A paired t-test is used to test the equality of means between provinces. A χ^2 test is used to test differences in proportions of categorical variables between provinces. Factors influencing production are production factors (land, labour, inputs), location, farm management, soil quality and household characteristics. As land input, the farm area that is actually used for cultivating food and cash crops is included. Two different sources of labour are distinguished, namely family labour (expressed in person units) and hired labour (expressed in paid wages). Family labour is measured as the number of adult family workers which we had to use as an admittedly imperfect proxy of time spent by family workers on the household farm because of lack of more detailed data. One other type of non-labour inputs is included: the sum of the expenditure on seed, chemicals and

[3] We thank an anonymous referee to put forward this idea.

[4] We acknowledge that calculating a transport cost per ton production would increase precision of the output measure. However, we lack reliable data on transport costs per ton.

[5] Gaulier et al. (2008)

[6] Beans, maize, manioc cassava, peanut, peas, potatoes, rice, soya beans and wheat are valued at opportunity cost. Banana, coffee, sweet potatoes and sorghum are valued at market value.

agricultural equipment.

Four different types of control variables are included: location, field characteristics, cropping pattern and household heterogeneity.

1. *Location*, which is considered by adding fixed community effects. As the capital of the Ngozi province is the third largest city in Burundi, access to assets and markets in nearby communities might be significantly higher than in more distant communities (in Muyinga).

2. *Field characteristics*. Indicators for field characteristics are land fragmentation, soil characteristics and use of soil improving farming technology. Land fragmentation is assessed by the Simpson index. This index varies from zero to one and is calculated by dividing the total sum of the different field surfaces squared by the square of total cropping area ($S = 1 - \sum s_i^2 / (\sum s_i)^2$). Farms with higher land fragmentation will demonstrate a higher Simpson index. Farmers were asked to assess the steepness of the plot and soil quality of each of their plots on a scale from one to four. In addition, information is gathered whether the plot is located in a march. This resulted in the calculation of three variables, one variable that indicates the share of the total cropping surface that has a steep slope, a second variable representing the share with good quality soil and a third variable indicating the share in the march. Two dummies are included to account for the use of chemicals and animal manure as soil improving farming techniques.

3. *Cropping pattern*. A mixed cropping pattern is quantified by the share of the total cropping surface used for either: staple crops, cash crops, banana or other crops.

4. *Household heterogeneity*. We control for household heterogeneity by including the following variables: age of the household head, the share of household income derived from off-farm activities and a dummy for extension (whether or not the household has been visited by an extension officer).

Finally, in our survey, we registered the Household Food Insecurity Access Scale or HFIAS

which was developed by the FANTA project of USAID (Coates and Bilinsky, 2007).[7] This HFIAS measures at household level several dimensions of food accessibility in the 30 days prior to the interview. To calculate the HFIAS the survey module assesses nine different dimensions of food accessibility, with for each dimension two specific questions: an occurrence question and a frequency-of-occurrence question.[8] The total of eighteen questions is then compiled into the HFIAS score. This is a continuous measure between 0 and 27 of the degree of food inaccessibility in the household in the past 30 days. Complementary information comes from the Household Food Insecurity Access Prevalence (HFIAP) score, which is a categorization of households into four levels by their set of responses: 1) food secure, 2) mildly food insecure, 3) moderately food insecure, 4) severely food insecure.

[7]The HFIAS method has been used in e.g. Frongillo and Nanama (2006), Knueppel et al. (2010), Gandure et al. (2010) and Becquey et al. (2010).

[8]Specifically, the nine questions deal with anxiety and uncertainty about the household food supply, insufficient quality of the food the household was able to obtain and insufficient food intake and its physical consequences.

Variables	Ngozi province		Muyinga province		Entire sample		Test
							t-test
Agricultural output (1,000BIF)	1132.83	(1158.07)	945.33	(1095.79)	1048.76	(1133.51)	2.07**
Market value output (1,000BIF)	1029.67	(1062.04)	787.60	(948.41)	921.13	(1019.01)	2.99**
Farm size (ha)	0.87	(1.44)	1.29	(1.89)	1.13	(1.66)	-2.26**
Farm size per person (ha/pers)	0.18	(0.24)	0.25	(0.35)	0.21	(0.29)	-2.68**
Size cultivated land (ha)	0.76	(1.1)	0.99	(1.45)	0.86	(0.52)	-2.12**
Yield (agr. output/cultivated land)	225.57	(230.13)	141.40	(124.76)	187.83	(194.67)	5.80**
Size cultivated land per person (ha/pers)	0.14	(0.20)	0.19	(0.29)	0.16	(0.25)	-2.59**
Family labour (nb)	2.74	(1.34)	2.51	(1.10)	2.64	(1.24)	2.30**
Labour cost (paid wage, 1,000BIF)	39.34	(13.66)	23.91	(100.77)	32.42	(118.35)	1.66**
Cost for seeds (1,000BIF)	20.46	(34.00)	17.62	(20.70)	19.18	(28.82)	1.28
Costs for chemicals (1,000BIF)	8.45	(20.56)	1.10	(5.98)	5.16	(16.19)	6.29**
Costs for agricultural material (1,000BIF)	4.47	(9.65)	3.76	(6.87)	4.15	(8.52)	1.02
Total cost production inputs (1,000BIF)	33.38	(48.38)	22.49	(25.00)	28.49	(39.98)	3.61**
Share in the marsh (%)	9.33	(12.28)	2.87	(6.29)	6.40	(10.54)	8.46**
Share under steep slope (%)	20.52	(29.85)	17.57	(29.59)	19.20	(29.75)	1.23
Share good quality soil (%)	49.51	(37.53)	46.49	(41.43)	48.15	(39.32)	0.94
Fragmentation index (0-1)	0.23	(0.14)	0.24	(0.14)	0.24	(0.14)	-0.51
Share staple crops (%)	52.51	(19.57)	61.88	(18.81)	56.71	(19.78)	-6.04**
Share coffee (%)	13.77	(13.62)	9.22	(10.71)	11.73	(12.60)	4.65**
Share banana (%)	20.78	(14.60)	18.05	(12.29)	19.55	(13.67)	2.53**
Share non-productive land use (%)	12.93	(17.27)	10.84	(17.02)	11.99	(17.18)	1.52
Age of hhhead (years)	41.36	(12.41)	40.01	(12.89)	40.75	(12.64)	1.32
Share income off-farm (%)	37.45	(3.59)	39.16	(32.04)	38.22	(32.33)	-0.65
HH Food Insecurity Access Scale (HFIAS)	13.66	(7.81)	14.60	(7.48)	14.08	(7.67)	-1.52
HFIAP (% severely food insecure)	0.68		0.73		0.70		1.51
HFIAP (% food secure)	0.08		0.06		0.07		1.69
							χ^2-test
Use of chemicals (% yes)	83		65		75		26.27**
Use of animal manure (% yes)	61		49		56		9.78**
Extension visit (% yes)	21		57		37		82.62**
Observations	342		278		620		

Significance levels : $*$: 5% $**$: 1% $***$: 0.1%

Note: t-values test for differences between the province means

Table 2.1: Descriptive statistics

2.3 Nonparametric regression approach

The empirical model is defined by a $n \times 1$ dependent scalar Y (total value of farm output)[9], a $n \times q$ multivariate regressor X (inputs) and an additive error ε.

$$Y_i = g(X_i) + \varepsilon_i \text{ , with } i = 1, ..., n. \tag{2.1}$$

This total value function can be estimated by imposing a parametric form. The vast majority of papers impose a Cobb-Douglas (CD) specification. Log output is defined as a linear function of the log of the q regressors, with additive error.

$$\ln Y_i = \alpha + \sum_{k=1}^{q} \beta_k \ln X_{ik} + \varepsilon_i. \tag{2.2}$$

It is from the assumed Cobb-Douglas specification of the production function that Assunção and Braido (2007) derive the yield approach that they prefer. The Cobb-Douglas specification implies that the factor elasticities are independent from scale and hence equal for all farms. However, if there are non-linearities or interactions in the true model, the empirical model is misspecified and the coefficients are inconsistent under a log-linear specification (Henderson and Kumbhakar, 2006). A flexible parametric alternative is the translog specification; quadratic effects and interaction effects are introduced in the empirical model.

$$\ln Y_i = \alpha + \sum_{k=1}^{q} \beta_k \ln X_{ik} + 0.5 \sum_{k=1}^{q} \sum_{l=1}^{q} \beta_{kl} \ln X_{ik} \ln X_{il} + \varepsilon_i. \tag{2.3}$$

In some cases, the translog specification can give economically unreasonable estimates, caused by (1) failure to capture all nonlinearities in the true model (Henderson and Kumbhakar, 2006), and (2) the high multicollinearity or low degrees of freedom as result of the inclusion of quadratic effects and interactions.

[9] We follow Assunção and Braido (2007) and Benjamin (1995) in regressing total value of farm output on the included covariates. As we value output in monetary terms, there is similarity with revenue or profit functions as discussed in e.g. Färe and Primont (1995) and Färe et al. (1990). However, as discussed, our sample consists predominantly of subsistence household farms with very low levels of marketing. Almost no revenues are actually generated, which makes it hard to presume we are estimating a revenue function. Further, we could not estimate a revenue or profit function because of the lack of individual farm-gate price data.

In order to allow for farm-specific input returns and yet avoiding to impose *'a priori'* a functional relationship between the total value of farm production and regressors, nonparametric approaches can be used[10]. In a nonparametric (generalized) kernel regression, $E[Y|X = x]$ is estimated by means of a localized regression (one could note it as $\hat{g}(x) = E[Y|\text{X close to x}]$).

Kernel weight functions are used to give more weight to observations near the observation point. Window widths impose the window of localization. If the window is large, the curve will be a smooth straight line. If the window width is small, non-linearities are allowed for and the curve becomes less smooth. It is intuitively clear and shown in literature that the choice of the weighting function is of far less importance than the choice of the window of localization - which we will discuss below.

We use kernel weights (l^c, l^u, l^o) with window widths $(\lambda^c, \lambda^u, \lambda^o)$ to specify the weight function for $x = [x^c, x^u, x^o]$, where x^c is a vector of continuous values, x^u is a vector of unordered discrete values, x^o is a vector of ordered discrete values. In particular, we specify a standard normal kernel function l^c to weight the continuous variable x^c_k (see (2.4)). An Aitchison and Aitken (1976) kernel l^u is specified to weight discrete unordered variable x^u_l with c_l categories and $\lambda^u_l \in [0, (c_l - 1)/c_l]$ (see (2.5)). To weight the ordered discrete value x^o_m, we use a Wang and van Ryzin (1981) kernel function with $\lambda^o_m \in [0, 1]$ (see (2.6)).

$$l^c\left(\frac{X^c_{ik} - x^c_k}{\lambda^c_k}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X^c_{ik} - x^c_k}{\lambda^c_k}\right)^2}. \tag{2.4}$$

$$l^u(X^u_{il}, x^u_l, \lambda^u_l) = \begin{cases} 1 - \lambda^u_l \text{ if } X^u_{il} = x^u_l, \\ \lambda^u_l/(c_l - 1) \text{ otherwise.} \end{cases} \tag{2.5}$$

$$l^o(X^o_{im}, x^o_m, \lambda^o_m) = \begin{cases} 1 \text{ if } X^o_{im} = x^o_m, \\ (\lambda^o_m)^{|X^o_{im} - x^o_m|} \text{ otherwise.} \end{cases} \tag{2.6}$$

[10] See Li and Racine (2007) for an extensive overview of the used kernel regression approach

To allow for a multivariate regression, we use - as is common practice - product kernels. The product kernel of x^c is $W_{\lambda^c}(X_i^c, x^c) = \prod_{k=1}^{q}(\lambda_k^c)^{-1} l^c((X_{ik}^c - x_k^c)/\lambda_k^c)$. For x^u, the product kernel is defined as $L_{\lambda^u}(X_i^u, x^u) = \prod_{l=1}^{r} l^u(X_{il}^u, x_l^u, \lambda_l^u)$. The product kernel of x^o is $L_{\lambda^o}(X_i^o, x^o) = \prod_{m=1}^{s} l^o(X_{im}^o, x_m^o, \lambda_m^o)$. All together, we can specify a Racine and Li (2004) generalized kernel function as $\mathcal{K}_\gamma(X_i, x) = W_{\lambda^c}(X_i^c, x^c) L_{\lambda^u}(X_i^u, x^u) L_{\lambda^o}(X_i^o, x^o)$, with $\gamma = (\lambda^c, \lambda^u, \lambda^o)$.

Two approaches were considered to estimate $E(Y|X = x)$. First, the Nadaraya-Watson estimator, which takes the kernel weighted average of the observed Y_i values and normalizes it by the sum of the kernel weighted averages (see (2.7)). This is the so called local-constant approach as it specifies a locally averaged constant value y for each observation point. It can be obtained as the solution of a in (2.8). Second, the local-linear estimator, which estimates a local linear relation for each observation point by obtaining a and b in (2.9). We opt for a local-linear estimator for two reasons. First, the main drawback of a local-constant estimator is that it can have a large bias near the boundary of support. The local-linear regression has better boundary properties than the local-constant regression (Hall et al., 2007). Second, if bandwidths are very large in a local-linear regression and there is thus no local weighting, we have the parametric least squares estimator. The least squares estimator can thus be seen as a special case of the local-linear estimator (Li and Racine, 2007, p. 83).

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i \mathcal{K}_\gamma(X_i, x)}{\sum_{i=1}^{n} \mathcal{K}_\gamma(X_i, x)}. \tag{2.7}$$

$$\min_a \sum_{i=1}^{n} (Y_i - a)^2 \mathcal{K}_\gamma(X_i, x). \tag{2.8}$$

$$\min_{\{a,b\}} \sum_{i=1}^{n} (Y_i - a - (X_i - x)'b)^2 \mathcal{K}_\gamma(X_i, x). \tag{2.9}$$

As discussed, the choice of multivariate bandwidth γ is of crucial importance. We opt for the often used data-driven approach that minimizes the asymptotic integrated mean squared error (AIMSE): the least-squares cross-validation approach as defined in (2.10).

$$CV(\gamma) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{g}_{-i}(X_i))^2 t(X_i). \tag{2.10}$$

where $\hat{g}_{-i}$ is the leave-one-out local-linear kernel estimator of $E(Y_i|X_i)$, and $0 \leq t(\cdot) \leq 1$ is a weight function that serves to avoid difficulties caused by dividing by 0 or by the slower convergence rate arising when X_i lies near the boundary of the support of X. Simulation results of Li and Racine (2004) show that cross-validated local-linear regressions indeed choose much larger bandwidths if the true relationship is linear.[11]

2.4 Empirical results

2.4.1 Description of the farming system related to farm size

The farming system in Burundi consists of small peasant landholdings (of generally less than 1 ha per family as illustrated in Figure 2.1), very small plots with double cropping, manual self-subsistence farming with little marketed surplus (Cochet, 2004). Crop production is done on both the hill side and in the drained marshes. Two distinct cropping systems were distinguished on each landholding. A first system consisted of separate plots cultivated with mixed crops (grains, pulses, tubers and coffee), and, a second system was based on banana production (see also Cochet (2004)). Several authors emphasize the importance of banana production in the current farming system (Rishirumuhirwa and Roose, 1998; Cochet, 2004). It seems as if the banana has over the years replaced cattle production which requires more land and other natural resources. The most important food crops produced and consumed in the study area were sweet potatoes, beans, cassava, banana and flour of maize (FAOSTAT, country profile, 2005). Except for banana and coffee, most farmers did not market their output and even when they did sell, it was mainly surplus sales of very small quantities.

The average farm size in our sample was 1.12ha however about 45% of the farms in the sample were smaller than 0.5ha. Farms were larger in Muyinga compared to the more densely populated Ngozi Province (see Table 2.1). The distribution of land over the sample was rather unequal. Moreover, compared to a previous survey we conducted in the same area in 1996 and with

[11] We opt for this approach over the AIC CV approach as the least-squares CV approach is more used in the literature and is faster to compute.

Rishirumuhirwa and Roose (1998), we find an increase of inequality in access to land, which resulted in an increased number of very small scale farms (smaller than 0.5ha). Furthermore farms were highly fragmented with on average more than eight plots on the hillside (collines), and one to two plots in the swamps (marsch). The relatively large farms in our sample are deliberately not excluded from the analysis as they may contain valuable information which can be studied separately with a nonparametric model.

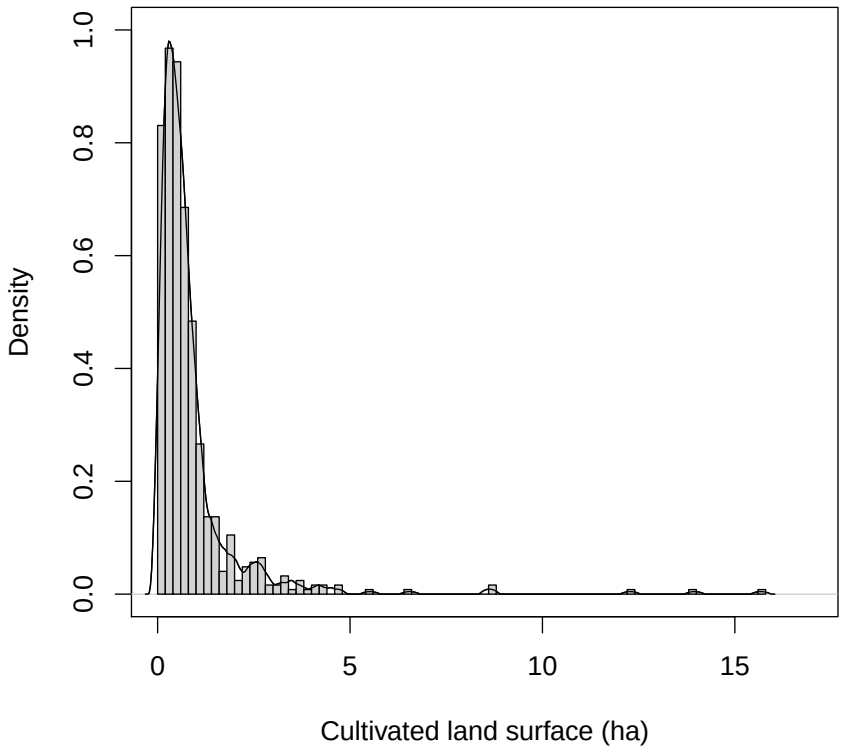


Figure 2.1: Density plot of farm sizes in the sample

In Table 2.2, we show descriptive statistics by quartile of production area. A one-way ANOVA F-test is used to test the equality of means between quartile groups. A χ^2 test is used to test differences in proportions of categorical variables between quartile groups. Yield as measured as the value of output per cultivated hectare is three times larger for the quartile of smallest farms

compared to the quartile of the largest farms. Uncontrolled for input use and environmental variables, the data hint at an inverse relationship between farm size as measured as cultivated hectares and farm productivity.

In addition, Table 2.2 suggests that farm size, production strategy, crop productivity and farm production may be related, although not all effects tend to go into the same direction. Large farms showed slightly different land use patterns compared to small farms. Larger farms tended to attribute a larger share of their total farm surface to other non-production activities such as forestry and fallow land whereas small farms used most of their land for staple food production rather intensively. However, the share of production area dedicated to cash crops, i.e. coffee production, did not significantly differ according to farm size quartiles. Small farms were using a larger proportion of the total production surface for banana production while larger farms used relatively more land for bean production (not detailed in the table). Farm proportions dedicated to other important crops in the area such as tubers and cereals did not differ between the land size quartiles and are therefore not reported. Crop diversification seems to be larger on larger farms, which is supposed to make them less prone to risks of crop failure compared to small less diversified farms.

Variables	First quartile		Second quartile		Third quartile		Fourth quartile		Test
									F-stat
Agricultural output (1,000BIF)	493.39	(383.40)	670.75	(406.48)	1017.89	(656.50)	2012.99	(1740.76)	75.69**
Market value output (1,000BIF)	429.55	(344.18)	601.84	(370.03)	902.33	(616.38)	1750.80	(1580.97)	68.06**
Farm size (ha)	0.22	(0.11)	0.56	(0.24)	0.91	(0.33)	2.82	(2.62)	119.46**
Farm size per person (ha/pers)	0.06	(0.05)	0.12	(0.08)	0.18	(0.12)	0.48	(0.48)	84.08**
Size cultivated land (ha)	0.16	(0.07)	0.41	(0.08)	0.71	(0.11)	2.2	(2.1)	111.95**
Size cultivated land per person(ha/pers)	0.05	(0.05)	0.13	(0.09)	0.18	(0.12)	0.48	(0.48)	84.08**
Yield (agr. output/cultivated land)	330.62	(309.32)	167.83	(112.28)	142.43	(90.59)	110.42	(82.45)	48.35**
Family labour (nb)	2.15	(0.68)	2.56	(1.17)	2.86	(1.39)	2.97	(1.43)	14.40**
Labour cost (paid wage, 1,000BIF)	7.34	(29.60)	10.22	(24.63)	22.04	(44.64)	90.08	(219.64)	18.19**
Seed cost (1,000BIF)	13.06	(21.05)	15.60	(17.99)	19.95	(20.23)	28.12	(44.11)	8.44**
Costs for chemicals (1,000BIF)	1.60	(0.05)	3.72	(11.21)	4.86	(14.90)	10.49	(25.32)	8.91**
Costs for material (1,000BIF)	2.59	(3.58)	3.57	(6.82)	4.32	(6.94)	6.13	(13.32)	4.86**
Total cost inputs (1,000BIF)	17.21	(23.44)	22.89	(29.08)	29.13	(32.54)	44.75	(59.54)	14.58**
Labour cost per ha (1,000BIF/ha)	28.19	(109.9)	17.80	(37.94)	23.45	(47.16)	34.78	(100.38)	1.25
Seed cost per ha (1,000BIF/ha)	72.40	(115.96)	28.71	(30.72)	23.46	(25.53)	14.61	(28.85)	25.87**
Costs chemicals per ha (1,000BIF/ha)	7.11	(19.23)	6.52	(20.07)	5.09	(16.11)	4.60	(12.17)	0.73
Costs material per ha (1,000BIF/ha)	14.86	(23.65)	6.36	(8.98)	4.99	(8.35)	2.76	(5.83)	21.32**
Total cost inputs per ha (1,000BIF/ha)	94.37	(131.49)	41.59	(46.03)	33.55	(34.80)	21.97	(36.38)	28.99**
Share in the marsh (%)	8.32	(13.83)	5.94	(8.87)	5.73	(9.48)	5.78	(9.05)	2.22.**
Share under steep slope (%)	18.65	(30.79)	20.30	(30.53)	16.37	(27.03)	21.48	(30.52)	0.86
Share good quality soil (%)	44.56	(38.74)	43.38	(38.34)	47.11	(38.92)	57.56	(40.04)	4.26**
Fragmentation index	0.30	(0.16)	0.23	(0.12)	0.23	(0.13)	0.19	(0.13)	17.27**
Share staple crops (%)	54.88	(21.27)	54.72	(19.41)	61.13	(18.63)	56.12	(19.20)	3.63**
Share coffee (%)	12.04	(14.45)	12.54	(11.61)	10.48	(11.28)	11.87	(12.84)	0.76
Share of banana (%)	22.53	(14.83)	19.67	(14.78)	18.04	(11.26)	17.99	(13.12)	3.82**
Share of non-productive land use (%)	10.55	(17.15)	13.07	(18.56)	10.35	(15.70)	14.01	(17.05)	1.76
Age of hhhead (years)	37.00	(11.37)	40.15	(12.39)	42.24	(13.46)	43.63	(12.37)	8.36**
Share income off-farm (%)	44.07	(33.71)	40.92	(33.65)	37.90	(31.12)	30.05	(29.26)	5.46**
HH Food Insecurity Access Scale (HFIAS)	17.75	(7.29)	15.18	(7.04)	13.47	(6.82)	9.92	(7.39)	32.86**
									χ^2-test
HFIAP (% severely food insecure)	0.81		0.74		0.71		0.54		29.30**
HFIAP (% food secure)	0.03		0.04		0.05		0.17		31.89**
Use of chemicals (% yes)	65.2		75.5		72.9		85.2		16.74**
Use of manure (% yes)	40.6		54.8		58.1		68.4		24.71**
Extension visit (% yes)	25.8		34.2		45.2		43.9		16.36**
Observations	155		155		155		155		

Significance levels : $*$: 5% $**$: 1% $***$: 0.1%

Table 2.2: Descriptive statistics by quartile of production area (N=620)

The allocation of labour seems to be closely related to farm size with larger farms allocating more family labour and spending more funds on hired labour. However, the level of labour per land unit was significantly higher for smaller farms as family labour per land unit was larger for small farms and wages paid for hired labour per land unit were not significantly different from larger farms.

Investments in agricultural production were measured by the expenditure on seed, agricultural material and chemicals. These investments increased significantly with increasing production

area. Smaller farms spent significantly more money per land unit on seed and agricultural material. The cost of seed per hectare of the lowest quartile is almost five times higher than that of the highest land size quartile. Fixed costs, the lack of seed reserves as result of severe food insecurity and more intensive use of material and seeds to make optimal use of the very small production area are possible explanations for this finding. Investments in chemicals such as fertilizer and pesticides were not different across the land size quartiles; these chemicals were used with the same, generally very low, intensity on both small and large farms. However the likelihood of using chemicals was larger on larger farms. On top of this, the likelihood of using specific soil improving techniques (manure, compost, mulching) was higher for the quartile with the largest farms.
These findings suggest differences in the production strategies related to differences in cropping area. These differences in crop production strategies might lead to different production outcomes and even more so to differences in farm productivity.

Symptomatic for the very poor livelihoods of the farm households in the study area, was their high level of food insecurity. Descriptive statistics show that only 7% of the households could be considered food secure (see Table 2.1). More than two thirds of all households interviewed were even labelled severely food insecure by the HFIAP score. These figures coincide with FAO data indicating that 68% of the total population is undernourished (FAO, 2009). Among these subsistence farmers, food (in)security is associated with cultivated land size. While half of the households in the highest production area quartile are categorized as severely food insecure, for the lowest quartile this is over 80% (see Table 2.2).

2.4.2 Results on farm size and productivity from the kernel estimations

The nonparametric approach makes no *'a priori'* assumptions on the functional relationship between the dependent variable and regressors. Using cross-validation, the trade-off between bias (for a given model, larger for a smooth, linear curve) and variance (larger for a wiggly, non-linear curve) is settled. As there is too few variation in family labour, it is inappropriate to consider this

as a continuous variable. Therefore, we define family labour as an ordered discrete variable.[12] In contrast to the parametric models, an ordered discrete variable can be included as one variable in a nonparametric model.

We illustrate the nonparametric results by showing directly in Figure 2.2 the estimated level of output as a function of the value of a respective independent variable, holding the other regressors equal to respectively the median for continuous variables or modus for (ordered) discrete variables. In addition, we show 95% bootstrap confidence intervals. A significantly increasing (resp. decreasing) curve illustrates a significant positive (resp. negative) effect of the regressor on agricultural production.[13] Rug plots on the x-axis illustrate the distribution of observations across the support of the variable.

The base model includes the size of land used for agricultural production, family labour, cost of hired labour, cost of inputs used, fixed community effects as independent variables. It also checks for the effects of field characteristics such as the steepness of the plots, perceived soil quality, share of land in marches, application of manure and chemical fertilizers, plot fragmentation (see Figure 2.2).[14] Steepness of the plots is particularly relevant for this hilly environment. The share of the farm located in the marches is of importance for the production of vegetables. The marches are drained and mostly used for vegetable production. Fragmentation is an important problem. The average number of plots on the farms in the sample is 6.6, with the largest quartile having on average eight plots.

The model shows significant effects of cultivated land and cost of hired labour. An increase in family labour did not significantly contribute to production. However, as the number of adult

[12]Results do no alter when we consider family labour as a continuous variable.

[13]The nonparametric model allows for interactions between all regressors. 3-D plots of estimated interactions between regressors are available on request.

[14]Preliminary analysis showed that inclusion of dummies for 'no labour cost' and 'no costs intermediary inputs' does not improve the localized model as the dummies are 'smoothed out' (bandwidth equal to 1). Table 2.3 in Appendix summarizes the used LSCV bandwidths. We make use of the 'np' package in R of Hayfield and Racine (2008).

family workers is an imperfect proxy of the effective family labour input, family labour should be considered as a control variable in this model. There is a clear non-linear relationship between hired labour and agricultural output. Further, we find a non-significant negative effect of steepness of the plots. Fragmentation has a significant non-linear effect. Perceived soil quality is found to be highly significant. Field characteristics are thus important determinants of agricultural production.

Because of the high correlation (0.44) between land surface and hired labour, the effects of the two variables are difficult to disentangle. The farm scale is therefore considered as a combination of both.[15] In Figure 2.3, we define the scale of the farm by the respective quantiles of hired labour and land surface used for production. A scale of 0 (resp. 1) means that the farm uses the minimum (resp. maximum) level of hired labour (larger than 0) and the minimum (resp. maximum) surface for production found in the data. As such, we can evaluate whether returns to scale depend on the scale (measured in terms of both labour and land) of the sampled farms.[16] Figure 2.3 illustrates that returns to scale of hired labour and land surface are a function of the scale of the farm. Relatively small farms are found to have returns to scale close to 0. Relatively large farms have returns to scale not far below 1. The assumption that returns to scale are not scale dependent - as imposed in the CD model and shown by the horizontal black line - is thus rejected at the 95% confidence interval. This finding of large heterogeneity in returns to scale is not sensitive for altering the controls from field characteristics to cropping pattern and household heterogeneity in the model or altering the valuation of agricultural output as shown in Appendix.[17]

[15]We do not consider the scale effect of intermediary inputs in this analysis because 1) the use of intermediary inputs is not highly correlated to land surface (correlation of 0.12) and 2) the effect of intermediary inputs is insignificant. It should be mentioned that both the physical and economic access to intermediary inputs are rather problematic in the area studied.

[16]We chose to focus on the heterogeneity in returns to scale, caused by differences in both land surface and hired labour. Note that the slopes in Figure 2.2 already illustrate possible heterogeneity in returns to scale, caused by either land surface or hired labour used, keeping everything else equal. We also made a 3-D plot representation of the dependency of returns to scale to variation in land surface and hired labour used. However, as it did not provide additional insight, it is not included but available upon request.

[17]We do not add variables concerning cropping pattern and household heterogeneity to the base model as non-

Based on this sample of small scale farms, we find clear indications of the size dependency of returns to scale. Figure 2.3 shows that scale elasticity varies from 0.2 for the smallest farms to 0.8 for the largest farms. As scale elasticity is significantly below 1 for every farm scale, we do not reject the occurrence of the Inverse Relationship.

parametric estimates can be unreliable, and thus overly insignificant, when the number of of continuous variables is too high (see e.g. Li and Racine (2007)).

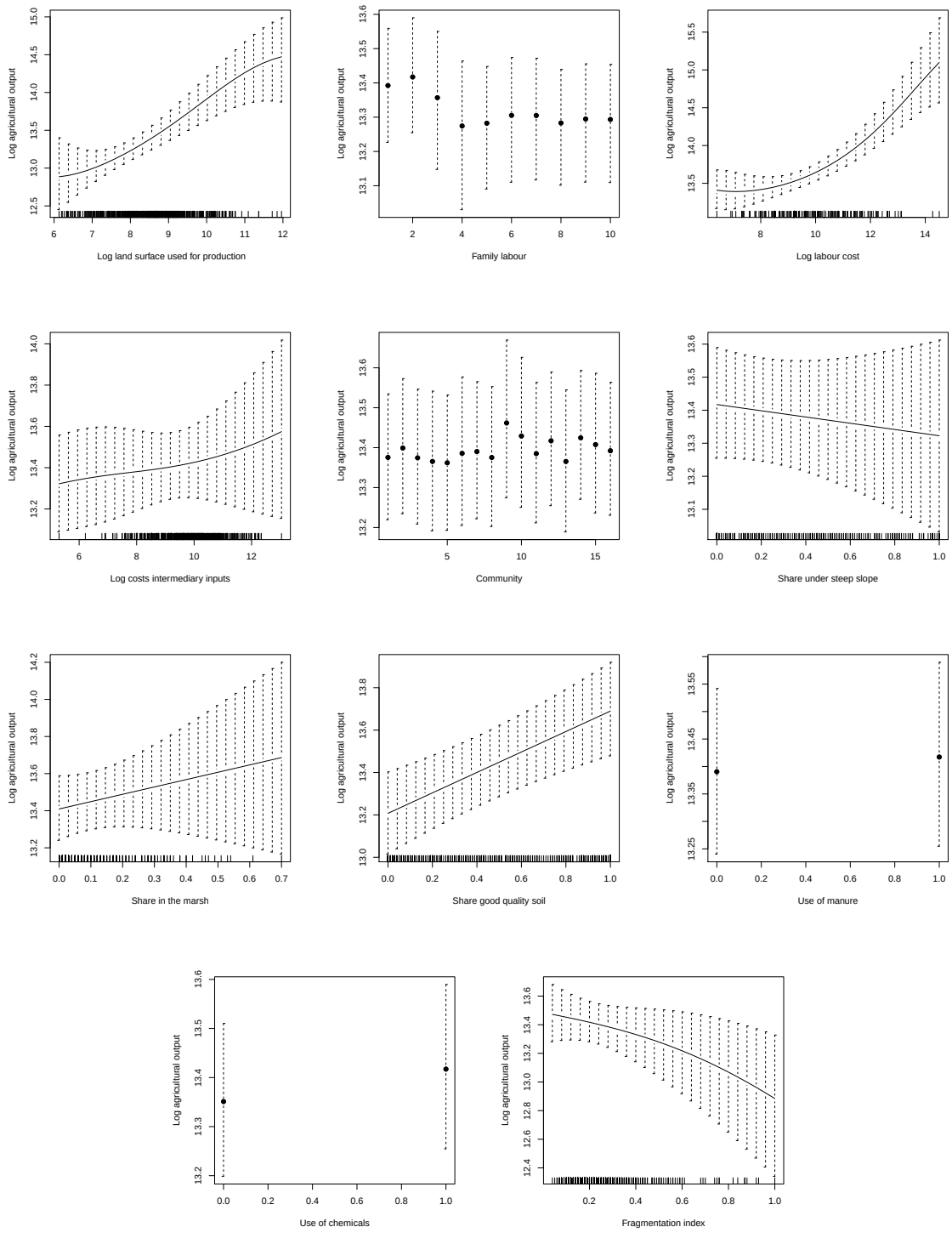


Figure 2.2: Base model

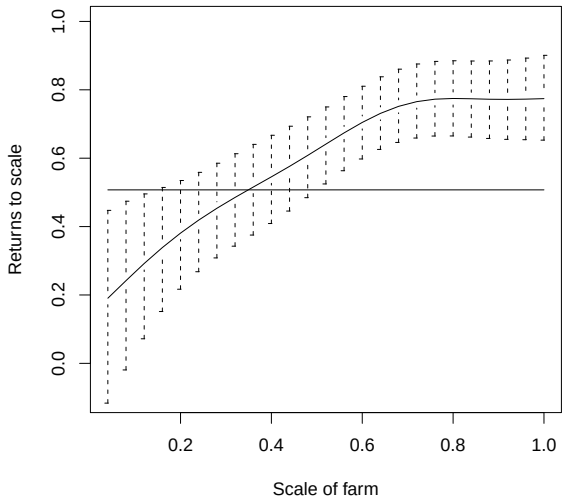


Figure 2.3: Returns to scale in function of scale of farm

2.4.3 Scale and food security

We showed in this study the large heterogeneity in scale returns in small-scale household farming. While a (marginal) increase in scale of 10% results for the smallest farm in only an increase of output of 2%, the same increase in scale would result in 6% increase of output for a middle-scale household farm and to 8% increase of output for a relatively larger farm. We were however not able to reject the inverse relationship by introducing heterogeneity in scale returns. From these results, one could argue that from an efficiency perspective, micro-farming is not a problem. As indicated in the Introduction, a number of hypotheses have been put forward in the literature with labour and factor market imperfections as the most important. As the sampled small scale farm holdings are characterized by self-subsistence farming with little marketed surplus, an additional relevant dimension in this debate may be the food insecurity issue, which we discussed in Section 2.3.1 to be severe for more than two thirds of the sampled farm households.[18] While it can be optimal from a productivity viewpoint to produce at small-scale, from a food security viewpoint, small-scale farming can have detrimental effects.

[18]We thank an anonymous referee to put forward this idea.

We estimated the relation between farm scale and food security by running our base model with log food security as dependent variable. We used the inverted Household Food Insecurity Access Scale or HFIAS (by calculating 28-HFIAS) such that a value of 1 indicates severe food insecurity and 28 food security. We add the variable off-farm income in this model to control for food security variation that is unrelated to household farm production.

Figure 2.4 shows a positive association between food security and farm scale (as measured in terms of hired labour and cultivated land size).[19] While larger farms perform better on the food security score, smaller ones are characterized by (severe) food insecurity. This indicates that from a micro-economic perspective, it is optimal to produce at a higher scale to ensure food security for the household members. In other words, we find a micro-level 'productivity-food security' trade-off. While it is optimal from a productive viewpoint to produce at very small-scale, there is a cut-off scale[20] of household farming activity under which severe food insecurity problems are expected as these households depend mainly on their farm production to supply sufficient food for its members. However, this result does not imply that food security and scale of farming are necessarily related at a macro-level. It is probable that the returns are used to ensure food security of the household members of the farm, while hired labourers are confronted with severe insecurity problems.

[19] Detailed results can be found in Appendix.

[20] As the used HFIAS score does not allow to define a cut-off between food security groups, we cannot actually estimate this cut-off scale.

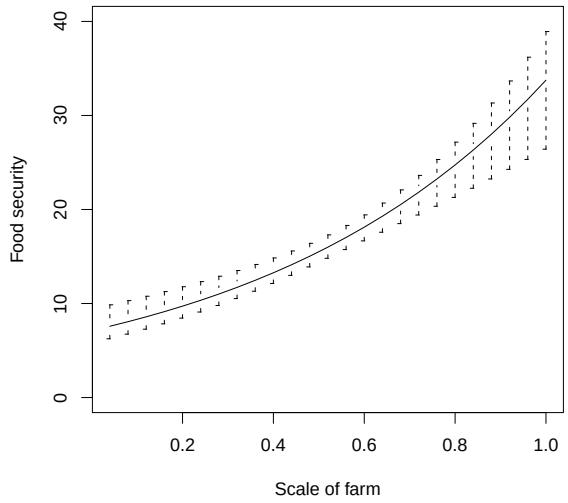


Figure 2.4: Food security and farm scale

2.5 Conclusions

The possibly inverse relationship between farm size and land productivity is one of the most persistent puzzles in development economics, even more so as many potential determinants have been put forward and tested without being able to provide an explanation. We study whether returns to scale are non-constant and whether this could contribute to the explanation of the occurrence of the IR. For this, we use data on small scale farm holdings in two Northern provinces of Burundi, which included information on the missing variables to which is referred in the literature. The sampled farms are characterized by considerable differences in output per hectare cultivated land between relatively smaller and larger household farms. In addition, there is heterogeneity in the use of inputs (small farms use more inputs per hectare) and field characteristics (small farms are more fragmented and have a lower share of plots with good soil quality).

We used a nonparametric kernel estimation of the production function (solved with a local-linear estimator) to allow for non-linearities and interaction effects. A base model was estimated controlling for inputs, location and field characteristics. Sensitivity tests were performed to control

for cropping pattern and household heterogeneity. We find a significant effect of land size and a non-linear effect of hired labour on agricultural output. In addition, field characteristics matter. Fragmentation and low perceived soil quality are associated with low agricultural productivity.

The nonparametric model confirms that farm size itself matters for the relationship between its size and productivity. Scale elasticity varies between 0.2 for the smallest farms and 0.8 for the largest farms and the assumption of a constant scale elasticity over the whole size range is rejected. In this sense we qualify the occurrence of the inverse relationship between farm size and land productivity, yet without fully explaining it. However, while small-scale farming is from an efficiency viewpoint unproblematic, production levels may be insufficient to guarantee food security of the household members. Smaller farms are in contrast to larger ones characterized by (severe) food insecurity.

2.6 Appendix

2.6.1 Sensitivity analysis

First, we control for cropping patterns (see Figure 2.5). The effects of cultivated land, costs for hired labour and intermediary inputs, and location are similar as for the base model. Farms with a larger share of banana cultivation are found to have a higher agricultural output. As the only cash crop, the share of coffee planted contributes positively to production. However, unobserved heterogeneity in among others the allocation of crops between fields of different quality could drive the found differences in productivity between cropping patterns. Therefore, we consider variables related to the cropping pattern as controls. Again, Figure 2.5(i) shows that returns to scale are scale dependent.

Second, we control for the effect of off-farm income in total household income, the access to extension services and the age of the head of the farm household (see Figure 2.6). We do not find significant effects of the three variables. The effect of farm size cultivated is not significant

in this model. Again, we find that returns to scale are dependent on the scale of the farm.

Third, we check whether the results are sensitive for altering the definition of agricultural output. Using the market value of all crops instead of the opportunity costs for food crops where possible did not alter any of the findings. Results available upon request.

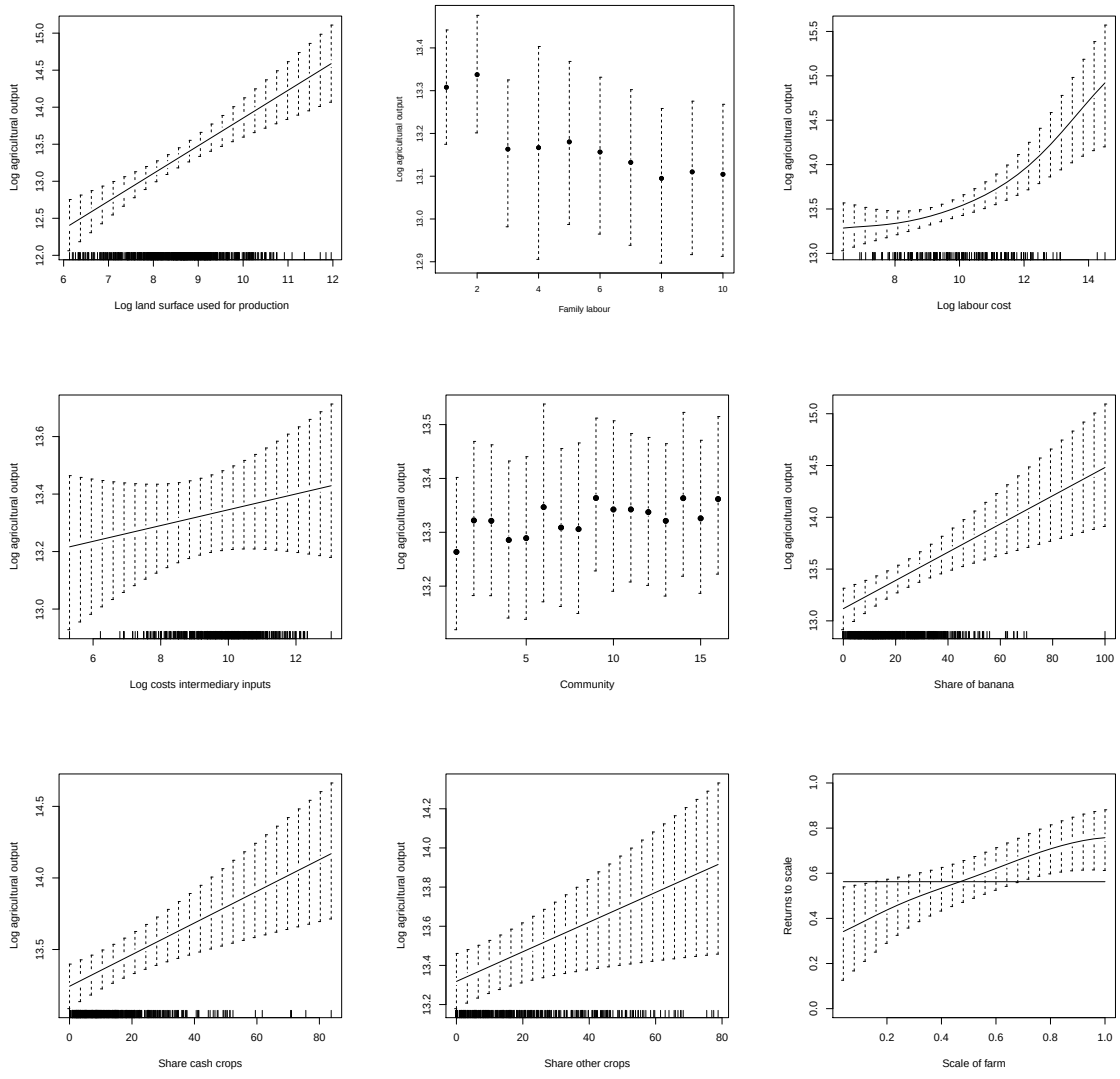


Figure 2.5: Model cropping pattern

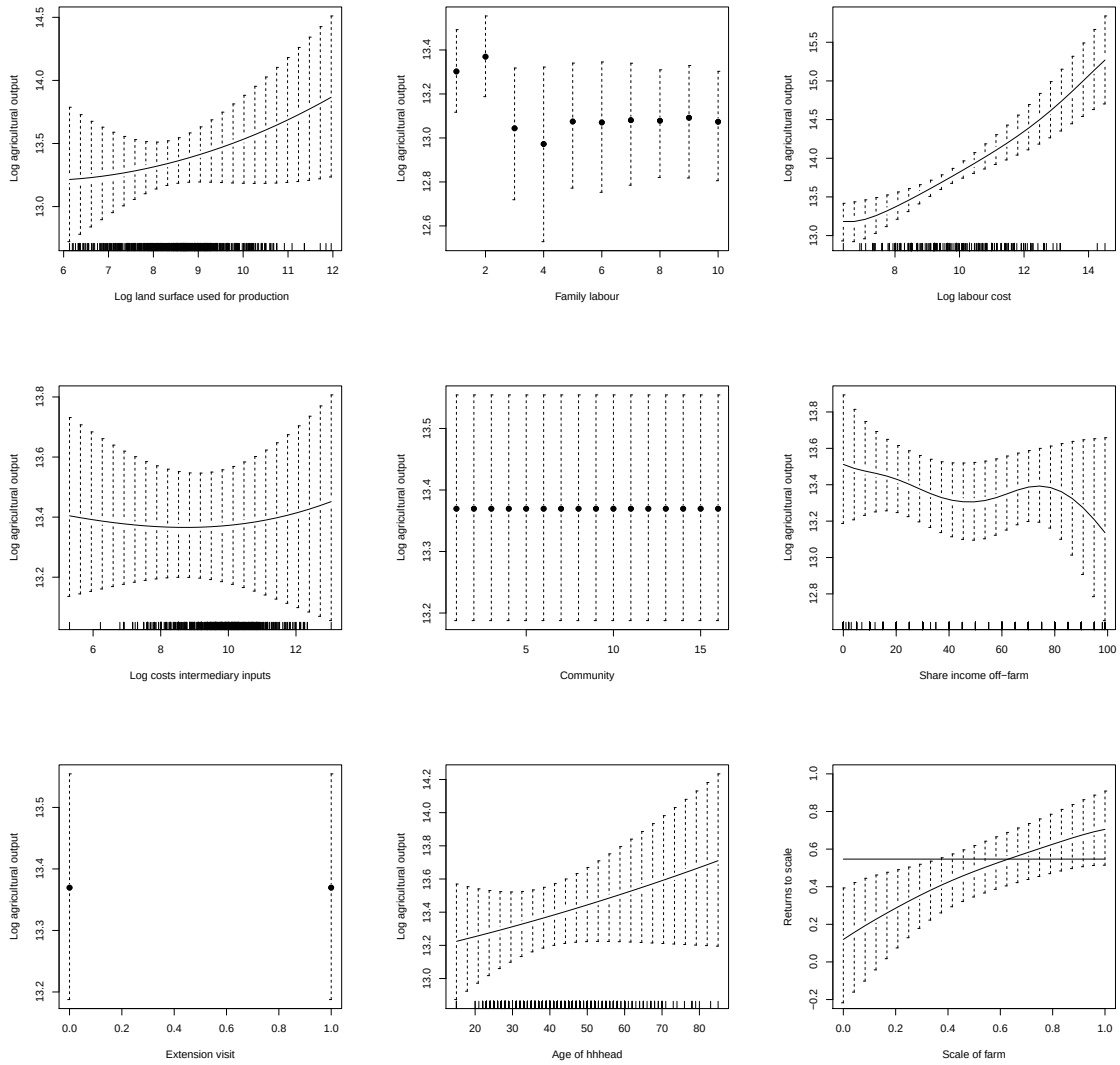


Figure 2.6: Model household heterogeneity

2.6.2 Food security

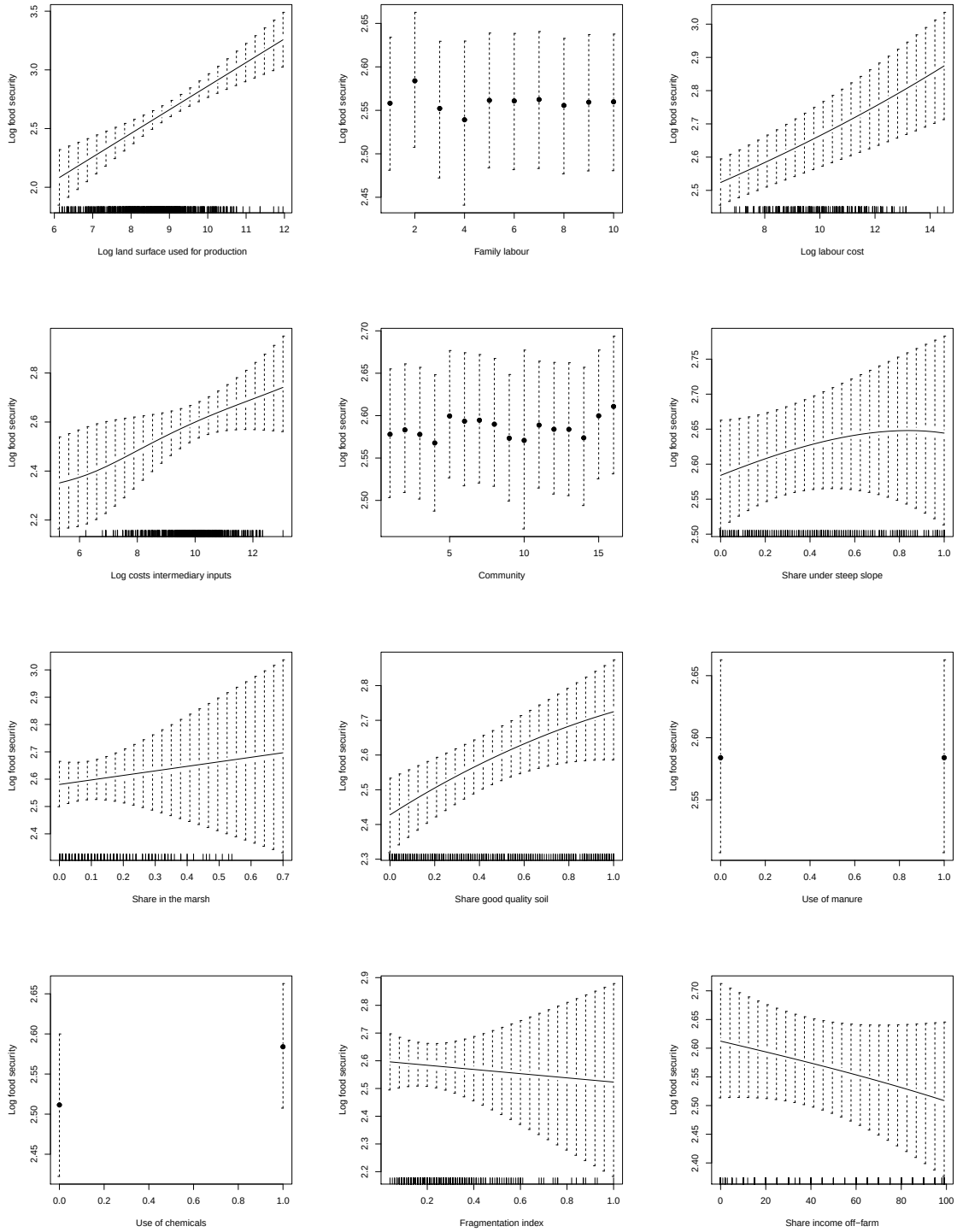


Figure 2.7: Model food security

2.6.3 Bandwidth sizes

	Model I	Model II	Model III	Model IV
Log cultivated land	1.23	$1.44e^5$	2.39	$4.57e^4$
Family labour	0.73	0.52	0.44	1.00
Log hired labour cost	1.75	1.52	1.89	8.46
Log costs intermediary inputs	3.38	$8.66e^6$	7.20	2.59
Community	0.86	0.83	0.94	0.86
Share under steep slope	$4.38e^5$			0.53
Share in the marsh	$5.69e^4$			$1.91e^5$
Share good quality soil	$1.04e^5$			0.58
Use of manure	0.43			0.50
Use of chemicals	0.27			0.18
Fragmentation index	0.48			$3.70e^4$
Share of banana		$4.16e^6$		
Share of cash crops		$1.12e^7$		
Share other crops		$9.25e^6$		
Share income-off-farm			15.13	83.00
Extension visit			0.50	
Age of hhhead			39.06	

Table 2.3: Bandwidths

References

Aitchison, J., Aitken, C. G. G., 1976. Multivariate binary discrimination by kernel method. Biometrika 63 (3), 413–420.

Assunção, J., Braido, L., 2007. Testing household-specific explanations for the inverse productivity relationship. American Journal of Agricultural Economics 89 (4), 980–990.

Assunção, J., Ghatak, M., 2003. Can unobserved heterogeneity in farmer ability explain the inverse relationship between farm size and productivity. Economics Letters 80, 189–194.

Barrett, C., 1996. On price risk and the inverse farm size-productivity relationship. Journal of Development Economics 51, 193–215.

Barrett, C., Bellemare, M., Hou, J., 2010. Reconsidering conventional explanations of the inverse productivity-size relationship. World Development 38, 88–97.

Becquey, E., Martin-Prevel, Y., Traissac, P., Dembele, B., Bambara, A., Delpeuch, F., 2010. The Household Food Insecurity Access Scale and an Index-Member Dietary Diversity Score contribute valid and complementary information on household food insecurity in an urban West-African setting. Journal of Nutrition 140 (12), 2233–2240.

Benjamin, D., 1995. Can unobserved land quality explain the inverse productivity relationship? Journal of Development Economics 46, 51–85.

Byiringiro, F., Reardon, T., 1996. Farm productivity in Rwanda: Effects of farm size, erosion, and soil conservation investments. Agricultural Economics 15, 127–136.

CIA, 2010. The world factbook.
URL https://www.cia.gov/library/publications/the-world-factbook/

Coates, J., S. A., Bilinsky, P., 2007. Household Food Insecurity Access Scale (HFIAS) for Measurment of Household Food Access: Indicator Guide (v.3). Food and Nutrition Technical Assistance Project, Academy for Educational Development, Washington D.C.

Cochet, H., 2004. Agrarian dynamics, population growth and resource management : The case of Burundi. GeoJournal 60, 111–122.

Collier, P., Dercon, S., 2009. African agriculture in 50 years: Smallholders in a rapidly changing world? In: FAO, UN Economic and Social Development Department.

Dercon, S., Krishnan, P., 1996. Income portfolios in rural Ethiopia and Tanzania: Choices and constraints. Journal of Development Studies 32, 850–875.

Eastwood, R., Lipton, M., Newell, A., 2010. Farm size. In: Pingali, P., Evenson, R. (Eds.), Handbook of Agricultural Economics. Vol. 4. Elsevier, pp. 3323–3397.

FAO, 2009. Food insecurity in the world, economic crises impacts and lessons learned. FAO (Food and Agriculture Organisation): Rome.

FAOSTAT, 2005. Country profile.
URL http://faostat.fao.org/

Feder, G., 1985. The relation between farm size and farm productivity : The role of family labor, supervision and credit constraints. Journal of Development Economics 18, 297–313.

Färe, R., Grosskopf, S., Lee, H., 1990. A nonparametric approach to expenditure-constrained profit maximization. American Journal of Agricultural Economics 72 (3), 574–581.

Färe, R., Primont, D., 1995. Multi-output production and duality: Theory and applications. Kluwer Academic Publishers.

Frongillo, E. A., Nanama, S., 2006. Development and validation of an experience-based measure of household food insecurity within and across seasons in northern Burkina Faso. Journal of Nutrition 136 (5), 1409–1419.

Gandure, S., Drimie, S., Faber, M., 2010. Food security indicators after humanitarian interventions including food aid in Zimbabwe. Food and Nutrition Bulletin 31 (4), 513–523.

Gaulier, G., Martin, J., Méjean, I., Zignago, S., 2008. International trade price indices. CEPII Working Paper Nr. 2008-10.

Hall, P., Li, Q., Racine, J. S., 2007. Nonparametric estimation of regression functions in the presence of irrelevant regressors. Review of Economics and Statistics 89 (4), 784–789.

Hayfield, T., Racine, J. S., 2008. Nonparametric econometrics: The np package. Journal of Statistical Software 27 (5), 1–32.

Heltberg, R., 1998. Rural market imperfections and the farm size-productivity relationship: Evidence from Pakistan. World Development 26, 1807–1826.

Henderson, D. J., Kumbhakar, S. C., 2006. Public and private capital productivity puzzle: A nonparametric approach. Southern Economic Journal 73 (1), 219–232.

Knueppel, D., Demment, M., Kaiser, L., 2010. Validation of the Household Food Insecurity Access Scale in rural Tanzania. Public Health Nutrition 13 (3), 360–367.

Lamb, R., 2003. Inverse productivity: Land quality, labor markets, and measurement error. Journal of Development Economics 71, 71–95.

Li, Q., Racine, J., 2004. Cross-validated local linear nonparametric regression. Statistica Sinica 14 (2), 485–512.

Li, Q., Racine, J., 2007. Nonparametric Econometrics: Theory and practice. Princeton University Press.

Lipton, M., 2010. From policy aims and small-farm characteristics to farm science needs. World development 10, 1399–1412.

Low, A., 1986. Agricultural Development in Southern Africa: Farm-household Economics and the Food Crisis. London: James Curry.

Platteau, J., 1996. The evolutionary theory of land rights as applied to Sub-Saharan Africa: A critical assessment. Development and Change 27, 29–86.

Racine, J., Li, Q., 2004. Nonparametric estimation of regression functions with both categorical and continuous data. Journal of Econometrics 119 (1), 99–130.

Rishirumuhirwa, T., Roose, E., 1998. The contribution of banana farming systems to sustainable land use in Burundi. Advances in GeoEcology 31, 1197–1204.

Schultz, T., 1964. Transforming Traditional Agriculture. Yale University Press: New Haven.

UNDP, 2010. Human Development Report 2010. The Real Wealth of Nations: Pathway to Human Development. United Nations Development Programme, New York, USA.

Wang, M. C., van Ryzin, J., 1981. A class of smooth estimators for discrete-distributions. Biometrika 68 (1), 301–309.

Wiggins, S., Kisten, J., Llambi, L., 2010. The future of small farms. World Development 38, 1341–1348.

World Bank, 2011. World development indicators.
URL `http:`
`www.worldbank.org`

3

School staff autonomy and educational performance: within school type evidence[1]

3.1 Introduction and related literature

A remarkable stylized fact of educational economics is that higher school resources do not necessarily yield higher pupil performance. Evidence for an overall large effect of school resource policies on pupil performance is largely missing (Hanushek, 2003; Wößmann, 2003). It cannot be excluded however that the effect of input-based policies on pupil performance is moderated by the incentive structure of school actors (Hanushek, 2003). Since the incentive structure of school

[1]This chapter is the result of joint work with Jean Hindriks (UCL), Glenn Rayp and Koen Schoors. Another version of this chapter appeared as CORE Discussion Paper 2010/82.

actors is embedded in school institutions, insight in the latter may be crucial to understand the relation between school resources and pupil performance.

There is accumulating cross-country evidence that getting incentives right by a combination of monitoring and autonomy is beneficial for educational efficiency (Wößmann, 2008). Since these studies compare institutions across countries the results may be biased because of an obvious omitted variable problem. Any other source of cross-country variation, like legal or cultural differences, may indeed bias the results. To identify the effect of school autonomy from within-country changes, Hanushek et al. (2011) use a panel estimation with country fixed effects on student level data from 42 countries from the Programme for International Student Assessment (PISA) 2000, 2003, 2006 and 2009 dataset. As school autonomy is considered as a country level average, within-country selection does not affect their estimates. Moreover, the use of country fixed effect estimation ensures that estimates are not affected by time-invariant country-specific heterogeneity. The authors find that the effect of school autonomy depends on the level of economic and educational development. In other words, in strong (weak) institutions, considerable academic gains (losses) are found from decentralized decision-making. In addition, the authors find larger gains from school autonomy if an adequate accountability system is in place such as central examination.

However, as secondary education is decentralized to the regional level in several countries (e.g. Belgium, Germany, Spain, United Kingdom), there is still regional level institutional variation that can affect the results from country fixed effects estimations.

Moreover, the assumption that institutional features such as the awareness of the importance of education and academic culture are time-invariant is hard in the specific setting. One example of time-varying institutional features is the change in institutional settings at country-level after the publication of dramatic scores for some countries in the first round of PISA (i.e., PISA 2000). This publication was the start of an "*an intense political debate that spread over almost all areas of the political and economic life, as the human capital acquired in a nation's schooling system is generally regarded as the most valuable resource of society*"(Ammermüller, 2004). Not without reason, the PISA 2000 publication was called the '*PISA shock*' in Germany.

In studies that use within-country variation of monitoring and autonomy, the problem of adequately controlling for (time-varying) institutional variation at country level is of course avoided. However, almost all student level within-country evidence of positive effects of school autonomy comes from studies comparing different types of schools (see e.g. the charter school literature, referenced in e.g. Abdulkadiroglu et al. (2011)). As noted by Hanushek et al. (2011), it is difficult to extract school autonomy effects from school type effects such as parental choice, quality of information or constraints on school location. Clark (2009) uses a different kind of variation, namely the variation in autonomy by the formation of a new type of public school with more autonomy (i.e., the '*Grant Maintained*' schools) in the UK between 1988 and 1997. The authors use a regression discontinuity design to compare public schools in which the vote to become a more autonomous '*grant maintained*' public school barely won with those in which the vote barely lost. Significant and persistent achievement gains are found for schools that opted to become '*grant maintained*' schools. Still, also in this study, school autonomy effects are not fully separated from other school type specific effects as variation in school autonomy is between school type.

The first contribution of this paper is that we employ within-country, *within school type* variation of autonomy on a dataset of pupil level performance, which gives us the degrees of freedom needed for statistical inference, and allows us to better isolate the school autonomy effects from school-type and country-specific effects.

This within school type variation comes from the particular structure of secondary education in Flanders, one of the three regions to which education is decentralized in Belgium. The constitutionally guaranteed freedom of education has resulted in four main school types (i.e., city schools, provincial schools, Flemish Community public schools, and (publicly-funded) privately operating catholic schools).[2] These educational entities have entrusted considerable school policy autonomy to non-profit school groups that can group several schools of the same type within the same city or region. School groups can determine school policy themselves or delegate op-

[2]About 75 percent of pupils are in catholic schools, nearly 25 percent in publicly organized schools and only a very small proportion of pupils are in non-catholic non-public schools.

erational autonomy to the school principals and teachers. We use variation in the amount of autonomy that trickles down to the lowest level, the school's direction and teachers as a strategy to identify the effect of school actor incentives on the relation between school resources and output.

However, selection issues and simultaneity problems can bias results from this kind of student level cross-sectional study. Therefore, we discuss the issues thoroughly and argue why our results are not driven by self-selection of students, reverse causality and the problem that more dynamic school teachers and principals will simultaneously boost educational performance and increase their operational autonomy, without there needs to be a direct effect of school autonomy on outcomes.

The second contribution of the paper is that we restrict ourselves to a narrow definition of autonomy. We only look into the effects of autonomy of principals and teachers, the local agents that, through their local informational advantage, are supposed to boost educational outcomes. This identification strategy brings our work closer to a clean test of the supposed effects of autonomy in a principal-agent framework (see below), where the government is the principal and the local school actors (school's direction and teachers) are the agents.[3]

The remainder of this chapter is structured as follows. In a second section, we discuss the theoretical background for the expected effect of school autonomy on educational performance. In section 3.3 and 3.4, we present the PISA data and the semiparametric multi-level analysis. In section 3.5, we discuss the results on the effect of school staff autonomy on educational achievement. Section 3.6 concludes.

[3]Hallinger et al. (1996) were the first to measure principals' activities in key dimensions of a school's instructional program and to relate these to student outcomes such as reading achievement. Wößmann (2003) is one of the first to look into the effect of individual teacher influence over teaching on student performance.

3.2 Theoretical background

The impact of school autonomy is linked to several strands of the literature. The decentralization of education may boost efficiency and productivity by eliminating unnecessary bureaucratic burdens (see Niskanen (1971) and Niskanen (1991), for seminal work on budget maximizing bureaucrats). School autonomy may help schools to overcome bureaucratic rigidity and in this way impact student performance positively (Bottani and Favre, 2001; Chubb and Moe, 1990). Entrusting the provision of education to local agents may also lead to more efficient provision because local agents will be closer and more responsive to student needs and preferences since students can 'vote with their feet' by changing school or even community. Tiebout (1956) shows that decentralized public good provision may, under certain conditions, yield the efficient provision of public goods like education. Hoxby (1999) confirms this Tiebout hypothesis for local school productivity under much less restrictive conditions. This suggests that the combination of decentralization and free school choice may indeed provide greater opportunities for local citizens and students to monitor and discipline the local agents that are responsible for educational policy, thereby creating greater efficiency and productivity.

If the decentralization of education is accompanied by public information on school performance, it may also be conducive to yardstick competition (see Shleifer (1985), and Besley and Case (1995)) among schools, in this way encouraging the adoption of more effective teaching methods and more efficient operational procedures. Card et al. (2010) for example recently find significant effects of enhanced competition on the test score gains of students in Canada in all studied school systems. They however also point at a possible negative effect of this yardstick competition. It cannot be excluded that "in more competitive markets teachers and principals spend more time and effort preparing for standardized tests, and less on other aspects of learning. If "test skills" have limited intellectual value, the effort devoted to competing over test outcomes is socially wasteful, and the higher test score gains observed in more competitive markets may be counterproductive" (Card et al. (2010), p. 29-30).

In weak institutional environments, decentralization may have some additional negative impli-

cations, like increased levels of uncoordinated rent-seeking and corruption as government structures become more complex and devoluted (Fan et al., 2009), increased coordination costs and slower institutional reform.

The most important negative consequence of increased autonomy may lie in a potential principal-agent problem (see Wößmann et al. (2007)). The government (principal) tries to improve cognitive skill creation by delegating responsibilities to schools (agents) that are assumed to have a local information advantage over the principal. A principal-agent problem appears when the interests of the government and the school diverge and information is asymmetric. Interests typically diverge for decisions that influence the financial position of the school or the workload for school actors. Budget formation and curriculum content are therefore policy areas with a high probability of divergence between the interests of the government and the school. In process and personnel decisions, little divergence of interest is expected. This principal-agent problem requires to put in place appropriate accounting systems. With effective accountability, autonomy is expected to enhance educational performance. Central examinations are a widely used accountability mechanism to align incentives between schools and the government (Wößmann et al., 2007), but other mechanisms can be used to attain this goal.

In Flanders, there are no central examinations, but inspection teams investigate on a regular basis whether the curriculum and teaching process are aimed at reaching the centrally imposed 'end goals' and whether budget formation is in accordance with the posed requirements. Benchmarking by parents is possible as the inspection reports are publicly available. In addition, freedom in budget formation is limited to additional funding, above the centrally imposed funding system. The size of these additional budgets is very small in comparison to the school budget. It mainly consists of revenues from student enterprises (such as a bakery in a bakery school) and donations by parents to finance e.g. school trips or school material (Poesen-Vandeputte and Bollens, 2008). Consequently, discretionary power of schools is limited on divisive issues like budget formation and curriculum development. We therefore expect that the principal-agent problem is limited in our case, and that the institution of school autonomy, through improved incentives for schools and teachers, affects resource-allocation decisions and ultimately the educational performance of

students positively.

It is worth noting that the positive effect of autonomy can also be supported by Oates' efficiency theorem (see e.g. Hindriks and Myles (2006, chapter 17) for an overview). Indeed under Oates' approach, autonomy is beneficial to better match local preferences and needs (the preference matching benefit) but could be detrimental in terms of lack of coordination and spillovers (the spillover costs). To determine if autonomy is beneficial it is necessary to compare the magnitude of the costs and benefits. It is easily seen that most items in school policy (such as budget formation, course content, teacher selection, disciplinary policies, student admission,...) display both preferences matching benefits and spillover costs. However the relative magnitude of these cost-benefits vary from one item to the other. Student and teacher selection is probably the one with highest spillover costs and the lowest preference matching benefit. It is therefore natural that we observe little school autonomy on such issues. On the other hand, the budget allocation presents low spillover costs and high preference matching benefits so autonomy is expected to produce better outcome than centralized decision making with uniform policy choice. In fact without spillover, decentralization is always superior. With spillover, decentralization can still dominate if there is sufficient difference across school in terms of needs and preferences. This argument relies heavily on the assumption that autonomy leads to better differentiation in educational policy to match local needs, and that the spillover effects are limited.

Last but certainly not least, autonomy is linked with intrinsic motivation. Human behaviour is driven by both intrinsic and extrinsic motivation. Both economic and psychological literature (i.e., Frey (1993), Frey (1994), Frey and OberholzerGee (1997) and Deci and Ryan (1985)) pinpoint the so called "*hidden cost of reward*", and cost of control by destroying the "*psychological contract*". Frey (1993) shows in a principal-agent framework that monitoring by the principal can be perceived by the agents as an indication of distrust, with lower work effort by the agents as result. Frey and OberholzerGee (1997) shows that extrinsic motivation (such as price incentives) has a crowding-out effect on intrinsic motivation. Differently put, price incentives or external intervention that is perceived to be controlling can reduce the feeling of living up to the civic duty and diminish altruistic behaviour. This implies that the intrinsic motivation of the school

staff and education quality is expected to be higher in schools where the school staff perceives to have considerable responsibilities.

3.3 Data

3.3.1 PISA 2006

We focus on the educational setting in Flanders, as its specific characteristics of education allow for a within school type analysis of school autonomy as we show below.[4] We use the PISA 2006 dataset. This because only for the sampled schools in 2006, we were able to group the schools in the four main school types in Flanders (as discussed, education organized by cities, provinces, the Flemish Community and publicly funded, privately operating entities (mainly catholic education)), whereas PISA only groups the schools in 'public' and 'publicly funded, privately operating schools'.

In 2006, the PISA survey was implemented in 57 countries. The main focus of PISA 2006 is on science, however all pupils are also requested to complete a standardized test on math, science and reading and fill out a survey with questions related to their family background, views on issues related to science, the environment, careers, learning time and teaching and learning approaches of science.

Tests are typically constructed to assess between 4,500 and 10,000 students of age 15 in each country. To sample the target population of 15-year old pupils that are at least in grade 7, PISA 2006 has implemented a two-stage stratified sample design. In stage 1, for each stratum[5], schools are sampled proportionally to size from a list of schools in the region (PPS sampling). The target was 150 schools in each region. In stage 2, 35 pupils are randomly drawn from a list of

[4]Since 1989, Belgian education is organized by the Flemish community, the French-speaking community and the German-speaking community. Hirtt (2007) argues that Flanders has the most effective accountability system of the three communities.

[5]A group of schools, formed to improve the precision of sample based estimates.

15-year old pupils in the school.[6] Final student weights are constructed to correct for varying selection probabilities of the students.[7] In PISA 2006 the plausible value approach is used to estimate the pupil performance in respectively mathematics, science and reading literacy. These plausible values are random values from the posterior distribution and cannot be aggregated at pupil level (OECD, 2005). Therefore, in what follows, we use the first plausible value component to estimate educational outcomes in math, science and reading at pupil level.[8] In Appendix, we discuss the interpretation of plausible values. A Balanced Repeated Replication (BRR) procedure with 80 replication estimates - described in OECD (2005)- is used to construct standard errors and to account for sampling variation (OECD, 2009).

Pupils in special education or part-time education are dropped from the sample. Pupils in private-funded schools or with missing values for some variables are also dropped from the sample. In addition, we do not take pupils in schools with less than 4 pupils per teacher into account as we expect that these schools have among others a different educational approach.[9] By this, the sample is reduced to 3603 observations. Sub-schools are defined to control for ability tracking in general, technical-arts and vocational education. A sub-school is defined as a unit that provides either general, technical-arts, or vocational education. When a school provides both general and technical or arts education, then the school is treated as two separate (sub-)schools. The sample consists of 126 schools and 245 sub-schools.

Table 3.1 shows descriptive statistics of the educational achievement of pupils in Flanders. Standardized test scores for math, science and reading are high in Flanders (PISA average is 500). But the high standard deviation of educational outcomes indicates that the inequality in individ-

[6] If the school has less than 35 pupils, all pupils are included in the sample.

[7] This occurs because certain subgroups that are over- or under-sampled, the information of school size at the time is not completely correct, school non-response, student non-response and the inclusion of trimming weights to ensure stable estimates.(OECD, 2009)

[8] As plausible values are random draws, the choice to take the first plausible value is arbitrary. Other plausible values could be taken as well.

[9] However, sensitivity tests available on request show that our results still hold when we do consider these schools with very few pupils per teacher as comparable and take them into the analysis.

ual test scores is also high.

Variable	Mean	St.Dev.
PISA 2006 Performance in math	560.3	(87.3)
PISA 2006 Performance in reading	543.0	(89.7)
PISA 2006 Performance in science	545.5	(82.0)
Difference PISA 2006 and PISA 2000 on reading	-10	(7.7)
Difference PISA 2006 and PISA 2003 on math	-10	(4.5)

Note: A SAS procedure for a Balanced Repeated Replication procedure with 80 replication estimates and 5 plausible values for each subject, described in OECD (2005), is used to construct the mean and standard error.
Source: OECD (2006) for last two rows

Table 3.1: Pupil performance

3.3.2 School staff autonomy

School autonomy is a rather vague concept. In our study we explicitly focus on autonomy of principals and teachers, using the data made available by PISA (2006). The PISA dataset among other things looks specifically into the roles that principals and teachers might play in educational decision-making and contains measures of centralization and decentralization for these different functions.

Table 3.2 illustrates how the level of school autonomy varies item-by-item in Flanders. In particular, the principal is asked who has the main responsibility for any specific item. The principal can tick multiple levels if there is joint decision making on a particular item. In line with Eurydice (2007) and Eurydice (2008), the PISA 2006 data show that Flanders is characterized by considerable school (group) autonomy in staffing, budget allocation and formation, assessment and discipline of pupils and textbook choice. Neither schools, nor intermediate government institutions have the autonomy to set the salaries of teaching and non-teaching staff. Although selection of students by schools is restricted to avoid exclusion of minority groups, the school's

direction remains largely in charge of approving the admission of pupils to the school.

However, these statistics need some qualification with respect to the sharing of decision responsibilities. As discussed in the Introduction, the Flemish government sets end goals that ought to be reached, but schools have considerable autonomy in how to reach these end goals. In practice, this amounts to a centrally imposed programme of basic courses, and considerable autonomy in the curricular content of optional subjects. As the end goals are detailed and well defined (see Hirtt (2007)), the principals can either interpret they have considerable autonomy (in reaching the end goals) or little autonomy (as there is little room to teach other things than the centrally imposed programme). In result, we note that principals tick both themselves and/or the educational authorities as decision makers in course content and courses offered. Although school groups or the school's principal have full authority to fire teachers, regulations strongly limit the possibility to fire a teacher, unless a serious fault is established.

We only consider in our analysis the variation in budget formation and allocation. As it is for these items that the school groups have not equally decentralized decision making to the school staff. Since the autonomy in budget formation is rather limited (as it needs to be in line with the centrally imposed funding system), our main focus is on budget allocation.

In short, Flanders is characterized by a combination of school staff empowerment in budgeting issues and accountability. We hypothesize that this combination of school staff empowerment in budgeting and effective accountability, after controlling for socio-economic and school-level institutional variation, has a positive effect on educational performance. As student and teacher selection into schools is non-random, extensive controls for student, teacher and school heterogeneity are needed.

Who has a considerable responsibility for the following tasks? (multiple ticks are allowed)	School's direction or teachers	Non-profit school groups	Regional or local education authorities	National education authorities
Selecting teachers for hire	120	51	4	3
Firing teachers	101	87	1	2
Establishing teachers' starting salaries	1	1	20	104
Determining teachers' salaries increases	1	1	19	106
Formulating the school budget	110	88	7	9
Deciding on budget allocations within the school	107	72	4	10
Establishing student disciplinary policies	125	33	6	9
Establishing student assessment policies	121	31	10	10
Approving students for admission to the school	117	16	11	24
Choosing which textbooks are used	125	6	1	2
Determining course content	88	7	28	64
Deciding which courses are offered	108	37	31	58

Table 3.2: School-level variation in perceived school autonomy, total of 126 schools

3.3.3 Control variables

Student characteristics To relate variation in outcomes to family background, we consider two socio-economic variables: socio-economic status and migration status (see Table 3.3). Family socio-economic status is estimated by PISA 2006 as a composite index of the Economic and Socio-Cultural Status (ESCS) of a pupil, derived from (1) the highest occupational status of each student's parents, (2) their highest educational level, and (3) a summary measure of household possessions. For the sampled students of all participating OECD countries, the mean is 0 and standard deviation is 1. The ESCS score shows substantial variation across pupils in Flanders.

For migration status, three proxies are used. First-generation immigrants and second-generation immigrants are respectively defined as pupils that are not born in Belgium and pupils that are born in Belgium, but are children of immigrants. Pupils that are first- or second-generation immigrant and do not speak an official Belgian language at home are grouped in a third variable. The proportion of non-native pupils is around 5 percent. 2 percent of the pupils in the sample do not speak an official Belgian language at home.

Pupils are tracked in the first year of secondary education in either general, technical-arts or vo-

cational education based on academic records. In our filtered sample, 50 percent of pupils are in general education (high track), 33 percent are in technical-arts education (middle track) and 17 percent in vocational education (low track). If a pupil has not reached the basic skills determined by the 'end goals' in a school year, grade repetition and re-orientation to lower tracks are used. In our sample, 79 percent of pupils are 'on time'.

School characteristics To control for school-level heterogeneity, we include controls for variation in educational resources, teacher shortage, class and school size, school type, social segregation, selectivity by schools and urbanization (see Table 3.3).

1. *Educational resources*. Schools receive funding and 'teaching hours' according to the number of pupils. Schools with more disadvantaged pupils receive additional resources ('GOK beleid'). On average, schools have a modest lack of educational resources (e.g. instructional material, labs) (the average is above the PISA 2006 average of 0).

2. *Teacher shortage*. In line with Rivkin et al. (2005) and Kane et al. (2006), preliminary analysis showed that the relation between formal teacher quality and educational performance is not significant. As almost all teachers in Flanders are certified, we dropped this variable. In contrast, shortage in educational personnel can have severe negative effects on the true teacher quality and teaching process in a school. Therefore, a negative sign is expected. Shortages of teachers in a specific area are too specific and for math and science, there is a systematic shortage in teachers.[10] Thus we focus on shortage of adequate teaching staff in areas other than math, science and reading.

3. *Class and school size*. The effect of school and class size on performance is difficult to disentangle from the selection effect (i.e. parents choose better schools that are consequently of larger size). Therefore, we use school size and student-teacher ratio to control for heterogeneity in class and school size.

[10] In preliminary analysis, we did not find effects of shortages in math, science or language teachers.

4. *School type*. We control for the 4 main school types in Flanders. Private-granted schools are only a negligible proportion of the school population. In our sample, 76 percent of pupils are in publicly funded, privately operating schools.

5. *Social segregation*. When schools organize different tracks, each school track is considered as a distinct sub-school. There is considerable variation in sub-school average ESCS, indicating social segregation between (sub-)schools. A quarter of the students are in sub-schools with an average ESCS below or equal to the OECD student average of 0, while the most elite sub-school groups students with an ESCS which is on average 64 percent higher than the OECD student standard deviation. As shown in Hindriks et al. (2010), there is less social segregation between school types than between school tracks.

6. *Selectivity by schools*. Selection by schools is officially not allowed within a track. However, Table 3.3 indicates selection on academic record or recommendation is frequently used.

7. *Urbanization*. Table 3.3 shows that 60 percent of the sampled pupils receive education in a town with 15,000 up to 100,000 inhabitants.

Variables related to school competition could be included. In Hanushek and Luque (2003), a significant positive effect is found of competition of private schools. Hoxby (2000) finds that Tiebout choice leads to better school performance in the US.[11] However, as we did not find any effect of a proxy for the number of competing schools in preliminary analysis, we dropped this variable to reduce the number of missing values.

Overall, we try to obtain insight into the relation between school autonomy -adequately measured - and educational performance, while controlling substantively for heterogeneity in student composition and institutional settings between schools. To further control for heterogeneity, we allow for non-linearities by the use of a flexible semiparametric econometric methodology.

[11] In Hoxby (2003), an overview of the economics of school choice can be found.

Variable	Students (%)	Schools	Private	Public
Autonomy in budget allocation	0.85	107	80	27
Autonomy in budget formation	0.88	110	81	29
Teacher shortage (other disciplines)	0.20	27	18	9
Selection by schools on academic record or recommendation	0.67	81	56	25
Achievement data used to evaluate principal	0.07	9	6	3
Organized by city	0.03	5		
Organized by province	0.04	5		
Organized by Flemish Community	0.17	24		
Subsidized private school	0.76	92		
Village ($<$ 3,000)	0.01	1	1	0
Small town (3,000 up to 15000)	0.29	35	29	6
Town (15,000 up to 100,000)	0.60	77	55	22
City (100,000 up to 1,000,000)	0.10	13	7	6
General education	0.50			
Technical-arts education	0.33			
Vocational education	0.17			
Female	0.48			
First-generation immigrant	0.03			
Second-generation immigrant	0.02			
Immigrant that speaks no off. Belgian language at home	0.02			
Not lagging behind	0.79			
Hours math per week: 0	0.04			
Hours math per week: less than 2	0.19			
Hours math per week: 2 up to 4	0.35			
Hours math per week: 4 up to 6	0.40			
Hours math per week: more than 6	0.03			
Hours Dutch per week: 0	0.04			
Hours Dutch per week: less than 2	0.22			
Hours Dutch per week: 2 up to 4	0.41			
Hours Dutch per week: 4 up to 6	0.31			
Hours Dutch per week: more than 6	0.01			
Hours science per week: 0	0.15			
Hours science per week: less than 2	0.31			
Hours science per week: 2 up to 4	0.32			
Hours science per week: 4 up to 6	0.13			
Hours science per week: more than 6	0.09			
Number of observations	3603	126	92	34

Table 3.3: Summary statistics, categorical variables

Variable	Mean	St.Dev.	Min.	25 perc.	Med.	75 perc.	Max
ESCS	0.29	0.85	-2.83	-0.32	0.27	0.94	2.99
Sub-school average ESCS	0.29	0.44	-1.30	0.02	0.32	0.63	1.64
School educational resources	0.10	0.84	-1.93	-0.38	0.09	0.46	2.14
Student-teacher ratio	9.27	2.38	4.11	7.56	9.20	11.30	14.04
School size	693.65	294.96	84	470	674	877	1712

Note: A SAS procedure for a Balanced Repeated Replication procedure with 80 replication estimates, described in OECD (2005), is used to construct the mean and standard error of the mean. The school educational resources index is a composite of the the quality of educational resources. It is composed from the principal's perception of shortage or inadequacy on 7 items of educational resources that can hinder instruction at school: 1) science laboratory equipment, 2) instructional materials (e.g. textbooks), 3) computers for instruction, 4) internet connectivity, 5) computer software for instruction, 6) library materials, 7) audio-visual resources.

Table 3.4: Summary statistics, continuous variables

3.4 Methodology

Educational settings are complex and heterogeneous. First, the largest part of the empirical data have a multilevel structure (pupils are nested within classes, classes within schools, schools within regions and school types, etc.). It is necessary to include this highly multilevel data structure into the empirical analysis to obtain unbiased estimates (Raudenbush and Bryk, 2002). This can be done by the use of a so called 'hierarchical' or 'mixed' model. This implies that the intercept - and in some models also the slopes - is allowed to randomly vary between groups. To estimate the effects of school-level institutional factors and family background on student achievement, a multilevel regression analysis is carried out where covariates are distributed at two levels: the students and schools. In an educational setting, unobserved school effects are expected from school-level disparities in e.g. the unobserved academic culture of school staff. As students are clustered in different schools, the assumption of independent noise is violated. It is thus necessary to include random school effects into the empirical analysis to obtain unbiased estimates.

Second, as result of the complex, heterogeneous nature of the data structure, imposing parametric assumptions on the relationship between educational inputs and output can lead to biased estimates if there is misspecification. As it is unclear how all variables affect educational performance, it is advisable to use a more flexible approach. Nonlinearities can be addressed in different ways. First, polynomial expansions can be considered. This would be easy to implement, but the risk of introducing multicollinearity is very high. Second, nonparametric approaches can be considered. Fully non-parametric approaches do not impose parametric assumptions on the functional form, but imply the so called 'curse of dimensionality' - that is that including a large amount of regressors dramatically slows down convergence speed - and involves practical difficulties to include random effects. To avoid the 'curse of dimensionality', we use a semiparametric additive mixed model approach.

We define pupil test scores of pupil i (with $i = 1,\ldots,n$) in school j (with $j = 1,\ldots,m$) as a function of socio-economic, institutional predictors and unobserved determinants such as innate ability and random noise at the pupil level $\varepsilon_{i,j}$. To allow for hierarchically clustered noise, we define θ_j as the random effect of school j. The semiparametric varying-intercept model is defined as:

$$\begin{aligned}
\text{PISA test score}_{i,j} =& \beta_0 + \beta_1 \text{School staff autonomy}_j \\
&+ \sum_{p=2}^{p=k} \beta_p \text{Student characteristic}_{p,i,j} \\
&+ \sum_{q=k+1}^{q=k+l} \beta_q \text{School characteristic}_{q,j} \\
&+ s_1(\text{ESCS}_{i,j}) + s_2(\text{Sub-school ESCS}_{i,j}) \\
&+ s_3(\text{School educational resources}_j) \\
&+ s_4(\text{Student-teacher ratio}_j) + s_5(\text{School size}_j) \\
&+ \theta_j + \varepsilon_{i,j}, \qquad (3.1)
\end{aligned}$$

where β_f, with $f = 1,\ldots,k+l$ are the *fixed parameters* of the categorical variables related to school staff autonomy, student characteristics and school characteristics and s_g, with $g = 1,\ldots,5$

are the *smooth functions* for the 5 additive continuous variables.[12]

Semiparametric regressions can be estimated by the use of kernel weights or by using piecewise polynomial functions - splines. Each approach has its own merits and drawbacks in a particular setting. We opt for a spline based approach as it is less cumbersome to use with large datasets and allows the inclusion of random effects. In particular, to smooth the continuous variables, we opt for the penalized splines (P-splines) approach of Eilers and Marx (1996), discussed in detail in Appendix.
The interest is not in the control variables per se. Therefore, if we find a non-linear effect of a smoothed variable, we only include information on the direction of influence. If the semi-parametric model pinpoints towards a linear relationship between educational performance and a specific continuous variable, we drop the smooth term and include the variable parametrically.

3.5 Empirical results

Table 3.5 shows six different models in which we measure the effect of autonomy of the school's staff in budget allocation on its educational performance in math. The focus is on math as Flanders is in this subject persistently ranked as a top performer (see e.g. De Meyer and Warlop (2010)). In addition, many teaching hours are devoted to this subject (see Table 3.3).

In a first model (i.e., Math I, first column of Table 3.5), we only control for student-level heterogeneity, school type and urbanization. Model Math I explains 55 percent of the variation of educational performance between pupils. We find an effect of migration status over and above the effect of the socio-economic status of pupils. The effect is amplified if the non-native pupil does not speak a Belgian language at home. We also find significant effects of school type and educational tracks. However, due to strong self-selection of pupils in school types and educational tracks, based on unobserved variables, a value-added approach is needed to obtain more reliable evidence on this matter. As expected, the control variables for grade repetition and teacher short-

[12]Multivariate smooths can also be introduced, but are not used in this analysis as multivariate tensor products of B-splines (see further) imply a dramatic loss of degrees of freedom.

age are significant.

Closer to our purpose, we find a clear positive effect of the autonomy in budget allocation on educational performance. Math performance is 13 points higher in schools with considerable autonomy in budgeting. As in all the following models, we do not find significant interaction between the effect of autonomy in budgeting and respectively school type or the average social position of pupils in a sub-school.

If we control additionally for school-level heterogeneity - by including teacher shortage, selectivity of the school, sub-school average ESCS, school educational resources, student-teacher ratio and school size- the relation of interest is still significantly positive, but is estimated to be only 6.73 points (see Model Math II). In line with Hindriks et al. (2010), we don't find an effect of the average sub-school ESCS after controlling for school type and educational track as social segregation occors largely between tracks and between school types. In contrast to what we would expect, we find a non-linear negative association between school educational resources and performance and a positive association of performance with the number of students per teacher. The direction of association between performance and school size is not clear as we find the relation to be locally positive and locally negative (i.e., wiggly).

Model	Math I	Math II	Math III	Math IV	Math V	Math VI
Plausible Value	PV1	PV1	PV1	PV1	PV1	PV1
(Intercept)	362.26***	362.62***	332.30***	327.36***	320.14***	310.20***
	(11.71)	(13.45)	(24.69)	(23.91)	(23.98)	(23.29)
Autonomy in budget allocation	13.43***	6.73*	10.68*	9.78°	10.84°	10.45°
	(2.80)	(2.92)	(5.36)	(5.25)	(5.59)	(5.38)
Privately operating school	18.48***	12.44***	14.78**			
	(2.40)	(2.69)	(4.72)			
General education track	117.06***	107.68***	103.07***	103.90***	100.99***	99.34***
	(3.57)	(4.72)	(5.54)	(5.50)	(5.74)	(5.72)
Technical-arts education track	65.90***	66.22***	65.74***	65.57***	63.93***	63.48***
	(3.28)	(3.49)	(3.64)	(3.62)	(3.79)	(3.78)
Gender (female=1)	-20.81***	-20.34***	-18.47***	-18.19***	-19.06***	-19.19***
	(1.97)	(1.97)	(2.07)	(2.07)	(2.15)	(2.15)
Small town	54.89***	74.50***	69.95***	68.55***	72.16***	80.29***
	(10.12)	(11.75)	(20.73)	(19.80)	(19.62)	(18.97)
Town	47.30***	67.37***	61.75**	63.29**	66.03***	73.34***
	(10.05)	(11.91)	(20.66)	(19.74)	(19.57)	(18.93)
City	43.63***	61.43***	56.00**	54.48**	57.71**	64.32**
	(10.50)	(12.36)	(21.32)	(20.36)	(20.16)	(19.62)
Lessons test subject, less than 2 hours	2.85	1.13	-2.12	-2.50	-1.22	-1.49
	(5.45)	(5.42)	(5.36)	(5.36)	(55.11)	(5.51)
Lessons test subject, 2 up to 4 hours	3.95	1.56	-2.98	-3.34	-3.05	-3.33
	(5.51)	(5.48)	(5.41)	(5.41)	(55.62)	(5.56)
Lessons test subject, 4 up to 6 hours	27.00***	24.82***	20.20***	19.83***	21.23***	21.11***
	(5.68)	(5.64)	(5.55)	(5.55)	(56.98)	(5.70)
Lessons test subject, 6 or more hours	33.66***	31.71***	24.78**	24.33**	23.93**	24.56**
	(8.02)	(7.94)	(7.80)	(7.80)	(7.97)	(8.00)
First-generation immigrant	-21.42**	-17.51*	-10.76	-10.86	-12.06°	-11.56
	(7.02)	(7.00)	(6.96)	(6.95)	(7.09)	(7.09)
Second-generation immigrant	-19.36**	-21.51**	-15.50*	-15.24*	-25.87**	-26.46**
	(7.25)	(7.47)	(7.34)	(7.33)	(8.26)	(8.25)
Immigrant, no off. Belgian language at home	-15.84°	-16.36°	-18.46*	-18.77*	-14.44	-13.96
	(8.86)	(8.89)	(8.70)	(8.70)	(9.02)	(9.02)
Not lagging behind	50.14***	49.34***	47.51***	47.41***	47.99***	47.62***
	(2.61)	(2.59)	(2.54)	(2.54)	(2.62)	(2.63)
Smoothed variables						
ESCS	6.05***	Pos.***	5.40***	5.40***	5.55***	5.51***
	(1.27)		(1.27)	(1.27)	(1.32)	(1.32)
Teacher shortage		-12.09***	-9.37°	-9.65*	-12.51*	-11.92*
		(2.68)	(4.81)	(4.69)	(4.95)	(4.78)
Selectivity of school		1.64	4.73	6.21	7.04°	5.72
		(2.19)	(4.01)	(3.936)	(4.07)	(3.94)
Average ESCS (sub-)school		2.33	2.14	2.21	3.21	4.47
		(3.83)	(5.06)	(4.97)	(5.05)	(5.00)
School educational resources		Neg.***	-2.73	Neg.	-4.57*	-3.70°
			(2.21)		(2.27)	(2.20)
Student-teacher ratio		Pos.***	3.28***	3.07***	3.44***	3.09***
			(0.96)	(0.92)	(0.94)	(0.91)
School size		Wiggly***	0.01	0.00	0.000	0.00
			(0.01)	(0.01)	(0.01)	(0.01)
Extracurricular science projects (including research)						9.08*
						(3.79)
Lectures and/or seminars on environmental topics						8.86*
						(4.24)
Student-level control variables	Yes	Yes	Yes	Yes	Yes	Yes
School-level control variables	No	Yes	Yes	Yes	Yes	Yes
Random school effects	No	No	Yes	Yes	Yes	Yes
Detailed school type FE	No	No	No	Yes	Yes	Yes
Control for reverse causality	No	No	No	No	Yes	Yes
Control for school dynamism	No	No	No	No	No	Yes
R^2 (adj.)	0.554	0.566	0.56	0.565	0.567	0.569
Obs.	3603	3603	3603	3603	3346	3318

Significance levels : ∘ : 10% : ∗ : 5% ∗∗ : 1% ∗∗∗ : 0.1%

Table 3.5: Effect of school autonomy in budget allocation, full model

In Model Math III, we include random school effects to control for unobserved school-level heterogeneity that is unrelated to the observed school-level variables. By taking this unobserved random school level heterogeneity into account, our model estimates it is optimal to employ no smoothing; we return to a fully parametric model. Results still hold.

As discussed, the PISA dataset groups pupils in 'Public schools' and 'Publicly financed, privately operating schools' to control for school type. However, as there are 3 public school types in Flanders (i.e., education organized by cities, provinces and the Flemish Community), we use anonymous data to control for the school type heterogeneity within public education. Model Math IV shows that the finding of a positive relation between educational performance and autonomy in budget allocation is not driven by school type differences. In particular, we find that variation in budgeting autonomy explains 0.1 standard deviation of pupil performance in math.

Overall, we extensively controlled for the heterogeneity in student population and institutions between schools. However, as we use cross-sectional pupil level data, estimates can be sensitive for influences of unobserved selection, reverse causality and simultaneity. However, we do not find indications that these issues drive our results.

First, on the self-selection problem, as discussed and argued above, we find no indications that (self-)selection of pupils and teachers into schools drives our results. As it is unobserved for students and teachers that consider to enter a school whether or not the school group or school's direction is in charge of budgeting issues, we do not expect that school choice by pupils and teachers is related to school staff empowerment in our specific setting.

Second, on the reverse causality problem, it is possible that the school direction in well-performing schools receive more autonomy. In other words, autonomy can be the result of performance, instead of the reverse. However, in the Flemish context, we do not expect this kind of reverse causality, since examination is not centralized. In result, it is not trivial for the non-profit school group to know how well a specific school performs. Furthermore, the PISA 2006 dataset indicates that in only a few schools (9 out of the 126 sampled schools), performance data (e.g. mean exam scores, average test results, rates of success in higher education) are used to evaluate

the school direction. In short, as evaluation of school policy is in most schools not based on performance, but more in terms of being in line with regulatory requirements, we do not expect this kind of reverse causality. Model Math V confirms this expectation, as we also find the significant positive relation between educational performance and school autonomy in budgeting for the subsample of schools in which there is no evaluation of the school's direction, based on achievement data.

Third, on the simultaneity problem, dynamic staff can in more autonomous schools use their freedom to among others organize lectures, plan field trips or guide extracurricular projects to raise academic standards. Differently put, autonomy gives the local staff the possibility to exert their dynamism. Estimation issues arise when the variation in the school staff empowerment in budgeting is related to unobserved variation in school dynamism. In specific, it is well possible that dynamic principals and teachers struggle for both more autonomy and higher educational achievement in the school, without there needs to be a causal relation between school autonomy and educational achievement .

To obtain insight whether this is the case, we first investigate whether our variation in school autonomy in budgeting is correlated with variation in school dynamism. To proxy school dynamism, we use PISA school-level questionnaire items on the promotion of science, opportunities to learn about environmental topics and preparation for further education. Schools that among others participate in competitions, organize lectures or guide projects in these fields are expected to also organize these extracurricular activities in other field and are labeled as dynamic. Table 3.6 shows there is little correlation between our proxies for school staff dynamism and school staff empowerment, indicating that our variation in staff autonomy is not simply reflecting the school staff dynamism. Therefore staff autonomy per se matters for educational performance

Further, if we include the proxies for school dynamism (separately or grouped) into our analysis, results for school autonomy do not change.[13] In particular, Model Math VI shows both a significant effect of two proxies for school dynamism (i.e., extracurricular science projects and lectures

[13]Results available on request.

and/or seminars on environmental topics) and school autonomy in budget allocation. No interactions were found. The found effect of school autonomy in budgeting is only significant at the 10 percent level when we consider the first plausible value of math as dependent variable. However, if we use the four other plausible values, the effect is significant at the 5 percent level, indicating a robust significant positive effect of autonomy in budget allocation on math performance (see Table 3.7).

	Autonomy in budget allocation	Autonomy in budget formation
Aut. in budget allocation	1.00	0.36
Aut. in budget formation	0.36	1.00
Science clubs	-0.05	0.08
Science fairs	0.10	0.08
Science competitions	0.10	0.11
Extracurricular science projects (including research)	0.04	0.00
Excursions and field trips related to science	-0.03	-0.13
Outdoor education related to environmental topics	0.04	-0.12
Trips to museums (related to env. topics)	-0.11	-0.09
Trips to science and/or technology centres	-0.05	0.06
Extracurricular environmental projects (including research)	0.15	0.10
Lectures and/or seminars on environmental topics (e.g. guest speakers)	0.02	0.06
Job fairs	0.07	-0.05
Lectures (at school) by business or industry representatives	-0.08	0.04
Visits to local businesses or industries	0.01	-0.07

Table 3.6: Correlation between proxies for school dynamism and school autonomy in budgeting

As discussed, autonomy in budget formation is limited to additional funding above a centrally imposed funding scheme. Table 3.8 shows a non-robust positive effect of school autonomy in budget formation. Once we control for the four main school types in Flanders, we do not find a (robust) significant effect of school staff empowerment in budget formation on math performance.

Finally, we test the sensitivity of our findings for test scores in different subjects. We ran Model VI for science and reading (see Table 3.8). For science, we find for both autonomy in budget allocation and budget formation a robust positive effect. As the additional budgets are very small in comparison to the school budget, we expect that the effect of budget formation is not mainly driven by an informational advantage of the agents. It is more plausible that higher perceived budget formation autonomy motivates the school staff (see Frey (1993)), with better education quality as result. For reading, we find the relation to be insignificant for both budget allocation and budget formation. A possible explanation is that autonomy in budgeting is not used to promote reading literacy. Another interpretation is that for reading, the unobserved heterogeneity is larger. For reading, only 52 percent of variation can be explained by Model Read VI, whereas Math VI and Science VI explain respectively 57 and 58 percent.

We have tested the robustness of the findings for different model specification. We have regressed a fully parametric model with random school effects, included variation in probability weights and clustering within strata. In addition, we have tested the effect on the results of including the 'dropped' variables. The findings remain robust to all such specification changes. Results are available upon request.

To test for a possible unequal effect of autonomy on educational performance, we compared the effects at the bottom and the top of test scores distribution by a quantile regression approach. Results in Appendix show no indications that school staff autonomy affects top and low performers differently.

Model	Math VI	Math VI	Math VI	Math VI
Plausible Value	PV2	PV3	PV4	PV5
Autonomy in budget allocation	13.99*	11.72*	11.04*	11.04*
	(5.54)	(5.54)	(5.54)	(5.56)
R^2 (adj.)	0.57	0.57	0.57	0.58
Autonomy in budget formation	11.63*	8.59	9.34	8.07
	(5.80)	(5.79)	(5.77)	(5.81)
R^2 (adj.)	0.57	0.57	0.57	0.58
Model	Science VI	Science VI	Science VI	Science VI
Plausible Value	PV2	PV3	PV4	PV5
Autonomy in budget allocation	12.82**	10.72*	12.20*	11.00*
	(4.87)	(4.82)	(4.94)	(4.88)
R^2 (adj.)	0.58	0.58	0.57	0.59
Autonomy in budget formation	17.01***	13.90**	15.79**	15.60**
	(4.93)	(4.92)	(5.02)	(4.94)
R^2 (adj.)	0.58	0.58	0.58	0.59
Model	Read VI	Read VI	Read VI	Read VI
Plausible Value	PV2	PV3	PV4	PV5
Autonomy in budget allocation	8.74	8.67	10.77	9.79
	(6.81)	(6.49)	(6.87)	(6.92)
R^2 (adj.)	0.52	0.53	0.51	0.52
Autonomy in budget formation	10.69	6.73	9.80	10.98
	(7.05)	(6.77)	(7.17)	(7.18)
R^2 (adj.)	0.52	0.53	0.51	0.53
Student-level control variables	Yes	Yes	Yes	Yes
School-level control variables	Yes	Yes	Yes	Yes
Random school effects	Yes	Yes	Yes	Yes
Detailed school type FE	Yes	Yes	Yes	Yes
Control for reverse causality	Yes	Yes	Yes	Yes
Control for school dynamism	Yes	Yes	Yes	Yes
Obs.	3318	3318	3318	3318

Significance levels : ∘ : 10% : ∗ : 5% ∗∗ : 1% ∗∗∗ : 0.1%

Table 3.7: Effect of school autonomy in budgeting, different plausible values

Model	Math I	Math II	Math III	Math IV	Math V	Math VI	Science VI	Read VI
Plausible Value	PV1	PV1	PV1	PV1	PV1	PV1	PV1	PV1
Autonomy in budget allocation	13.43***	6.73*	10.68*	9.78°	10.84°	10.45°	10.91*	11.12
	(2.80)	(2.92)	(5.36)	(5.25)	(5.59)	(5.38)	(4.78)	(6.97)
R^2 (adj.)	0.55	0.57	0.56	0.57	0.57	0.57	0.58	0.52
Autonomy in budget formation	8.09**	7.26*	10.21°	7.32	7.63	7.02	14.62**	10.77
	(3.00)	(3.11)	(5.73)	(5.60)	(5.82)	(5.61)	(4.86)	(7.28)
R^2 (adj.)	0.55	0.57	0.56	0.56	0.57	0.57	0.58	0.52
Student-level control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School-level control variables	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Random school effects	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Detailed school type FE	No	No	No	Yes	Yes	Yes	Yes	Yes
Control for reverse causality	No	No	No	No	Yes	Yes	Yes	Yes
Control for school dynamism	No	No	No	No	No	Yes	Yes	Yes
Obs.	3603	3603	3603	3603	3346	3318	3318	3318

Significance levels : ∘ : 10% : ∗ : 5% ∗∗ : 1% ∗∗∗ : 0.1%

Table 3.8: Effect of school autonomy in budgeting

3.6 Conclusion

In this paper we have tested whether a combination of accountability and autonomy in education provision can improve educational performance as measured by PISA test scores. There is ample cross-country and cross-school type (e.g. charter school literature) evidence for this. But these studies cannot separate the effects of school autonomy from (time-varying) country-specific or school type specific effects.

We test this conjecture on the PISA-dataset for Flanders, where there is substantial variation within school-type in autonomy and other institutional settings. In Flanders, the government delegates a lot of budgeting responsibilities to school groups. There is variation in the extent this budgeting authority is further delegated to the school staff (the school's direction and teachers). Higher school staff empowerment should lead to better use of local information and better performance, if the central government can align incentives properly. In Flanders, there are no central examinations, but inspection teams investigate on a regular basis whether the curriculum and

teaching process are aimed at reaching the centrally imposed 'end goals', and whether budget formation is in keeping with the posed requirements. In addition, freedom in budget formation is limited to additional funding, above the centrally imposed funding system. This ensures that information asymmetries are not misused by local staff and so that autonomy should improve performance. Therefore Flanders is a very good testing ground for the theory that the institution of school autonomy, through improved incentives for schools and teachers, will affect resource-allocation decisions and ultimately the educational performance of students positively.

Our findings support this hypothesis. While including a large set of student-level and school-level controls, we find indeed that local staff empowerment clearly and significantly boosts educational outcomes. Results are robust for controlling for reverse causality or variation in school dynamism.

Overall, we do not provide a natural experiment, but provide smoking gun evidence of a significant effect of school staff empowerment in budgeting on educational performance.

3.7 Appendix

3.7.1 Plausible values

PISA 2006 uses a plausible values approach to come to student population estimates of knowledge and skills in math, science and reading literacy. A plausible value approach was developed for and used in the 1983-84 US National Assessment of Educational Progress (NAEP). Thereafter, it is used in among others the TIMSS and PISA survey. A detailed discussion of the plausible value technique is given in Wu (2005) and OECD (2005, chapter 6). We briefly discuss the interpretation of plausible values.

The main problem in cognitive testing is that the latent pupil skills and knowledge is unobserved. Testing for skills by e.g. a PISA questionnaire involves thus measurement error above the sampling error. In social sciences, the measurement error is expected to be substantial, first, as result of the broadness of the concept that is measured and, second, because tested pupils may be affected by day-to-day (mental and physical) variation and conditions under which the test

occurs (OECD, 2005). In result, the measurement error depends on the precision of the test and on pupil-level characteristics. Population statistics will be biased if the measurement error is not taken into account. To construct unbiased population estimates, first, the distribution of student ability is estimated, using the (discontinuous) test items and background variables. Second, random draws are taken from this so called '*posterior*' distribution of student skills. Plausible values are thus multiple random draws from the unobservable latent student achievement. The standard error between the plausible values gives an indication of the magnitude of pupil-level measurement error. As discussed in OECD (2005), a priori averaging plausible values to conduct pupil-level inference leads to biased estimates. One should use the plausible values to do the regressions. Afterwards, one can take the average of the coefficients if wanted.

3.7.2 The P-splines approach of Eilers and Marx (1996)

A large methodological literature has focused on the issue how to represent smooth functions and to choose the smoothness of these functions (Wood, 2006). The popular backfitting approach of Hastie and Tibshirani (1990) has the benefit that multiple smooth terms can be included, with the cost that the model selection (= selection of number of smooths) can be quite cumbersome (Wood and Augustin, 2002). The alternative approach of Gu and Wahba (1991) has solved the model selection problem but at a high computational cost limiting its use. The regression spline approach is a computationally efficient approach to estimate a semiparametric additive model with integrated model selection (see among others Eilers and Marx (1996), Marx and Eilers (1998), Wahba (1980) and Wahba (1990)). We use this approach as implemented in the *mgcv* package in R with automatic and integrated smoothing parameter selection (see Wood (2006)). The spline approaches are not suited to include categorical variables, and so we include the those variables parametrically. We thus have a semiparametric partially linear mixed model.

The smooth function of a spline approach is a weighted sum of a basis of r overlapping splines. Figure 3.1(a) illustrates a cubic spline with local support (B-spline)[14]. By altering the weight of

[14]A univariate B-spline of degree q smoothly joins q+1 polynomial pieces of degree q at q interior knots in the local support. The local support implies that outside the boundaries, the value is zero.

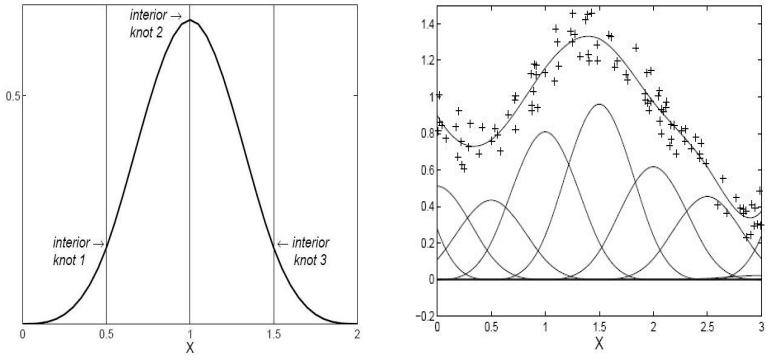
the splines by weight parameter α_j, with $j = 1, \ldots, r$ on usually evenly spaced knots in function of minimization of the squared error, we obtain a flexible nonparametric smooth - as shown in Figure 3.1(b). Formally, the smooth function $\hat{s}(x)_{(\alpha)_i}$ can be represented as the sum of r overlapping basis functions, multiplied by the respective basis parameters α_j, with $j = 1, \ldots, r$.

$$\hat{s}(x)_{(\alpha)_i} = \sum_{j=1}^{r} \alpha_j B_j(x), \text{ such that } \forall x, \sum_{j=1}^{r} B_j(x) = 1. \tag{3.2}$$

To estimate a regression via P-splines, α is estimated by minimizing the squared error (known as the L_2 norm) with inclusion of a penalty on wiggliness for each smooth function to avoid oversmoothing. Usually, the second order differences are penalized (d=2), however other penalty structures are also possible. As in Bollaerts et al. (2006), we define the L_2 norm as follows:

$$L_2 = \sum_{i=1}^{m} (y_i - \hat{s}(x)_{(\alpha)_i})^2 + \lambda \sum_{j=d+1}^{r} (\triangle^d \alpha_j)^2, \tag{3.3}$$

with $\triangle^d \alpha_j$ being the d^{th} order differences, that is $\triangle^d \alpha_j = \triangle^1(\triangle^{d-1}\alpha_j)$ with $\triangle^1 \alpha_j = \alpha_j - \alpha_{j-1}$ and with λ a non-negative smoothness parameter. It can easily be shown that the L_2 can be extended to include simultaneously a parametric part, random effects and univariate smooths in an additive approach.



(a) Univariate cubic B-spline basis (b) Smooth via B-splines

Figure 3.1: B-splines, source: Bollaerts (2009)

3.7.3 Effects at the top and bottom

To test for a possible unequal effect of autonomy on educational performance, we compare the effects at the bottom and the top of test scores distribution. For this, we estimate a quantile regression as initiated in the seminal work of Koenker and Bassett (1978). This approach allows a more complete picture of the conditional distribution of pupil performance. In this approach the conditional αth quantile ($\alpha \in (0,1)$) is defined as the test threshold such that α percent of the pupils of the reference group perform worse. For example, in the socio-economic status x, a quarter of the pupils performs worse than the score threshold $q_{0.25}(x)$. It is common practice to use a so called 'check function approach' to estimate a quantile regression via minimization of weighted absolute deviations from the fit.

$$L_1 = \sum_{i=1}^{n} \rho_\theta(y_i - \mathbf{x}_i\hat{\beta}), \tag{3.4}$$

with $\mathbf{x}$ a vector of regressors, β a vector of coefficients and with check function ρ_θ being defined as

$$\rho_\theta(\tau) = \begin{cases} \theta\tau & \text{if } \tau \geq 0 \text{ (resp. } \tau \leq 0) \\ (\theta-1)\tau & \text{otherwise,} \end{cases}$$

with τ being defined as $y_i - \mathbf{x}_i\hat{\beta}$. Weight factor θ indicates how positive and negative values of τ are weighted. If $\theta = 0.5$, positive and negative values are equally weighted and the median is estimated. If $\theta = 0.75$, positive values of τ receive a weight that is three times higher than the weight of negative values; the third quartile is estimated.

However, a drawback of a quantile regression approach is the lack of a consensus on how to include random school effects in the model. As such, the advantage of a more complete picture of the conditional distribution of pupil performance comes at the cost that we cannot control for random school effects.

We opt for the parametric quantile regression approach, as implemented in the package 'quantreg' in R of Koenker (2011). We estimate 20 quantile regressions of model Math VI (without random school effects) with θ between 0.05 and 0.95.

The results in Figure 3.2 and 3.3 show the relation between the conditional quantile (x-axis)

and the estimated coefficient (y-axis) for model Math VI with autonomy in budget allocation included. The shaded area denotes the 90% confidence interval. The results show that the effects of family background, social segregation and shortage of adequate teaching personnel are larger for pupils in the lower end of the distribution of math performance. Further, the quantile regressions show robustness of our findings. School staff empowerment has a significant positive effect on pupil performance. Lastly, we find no significant differences in the effect of school autonomy in budgeting over the conditional distribution of pupil performance.

In line with the semiparametric analysis, results for budget formation are less pronounced (see Figure 3.4).

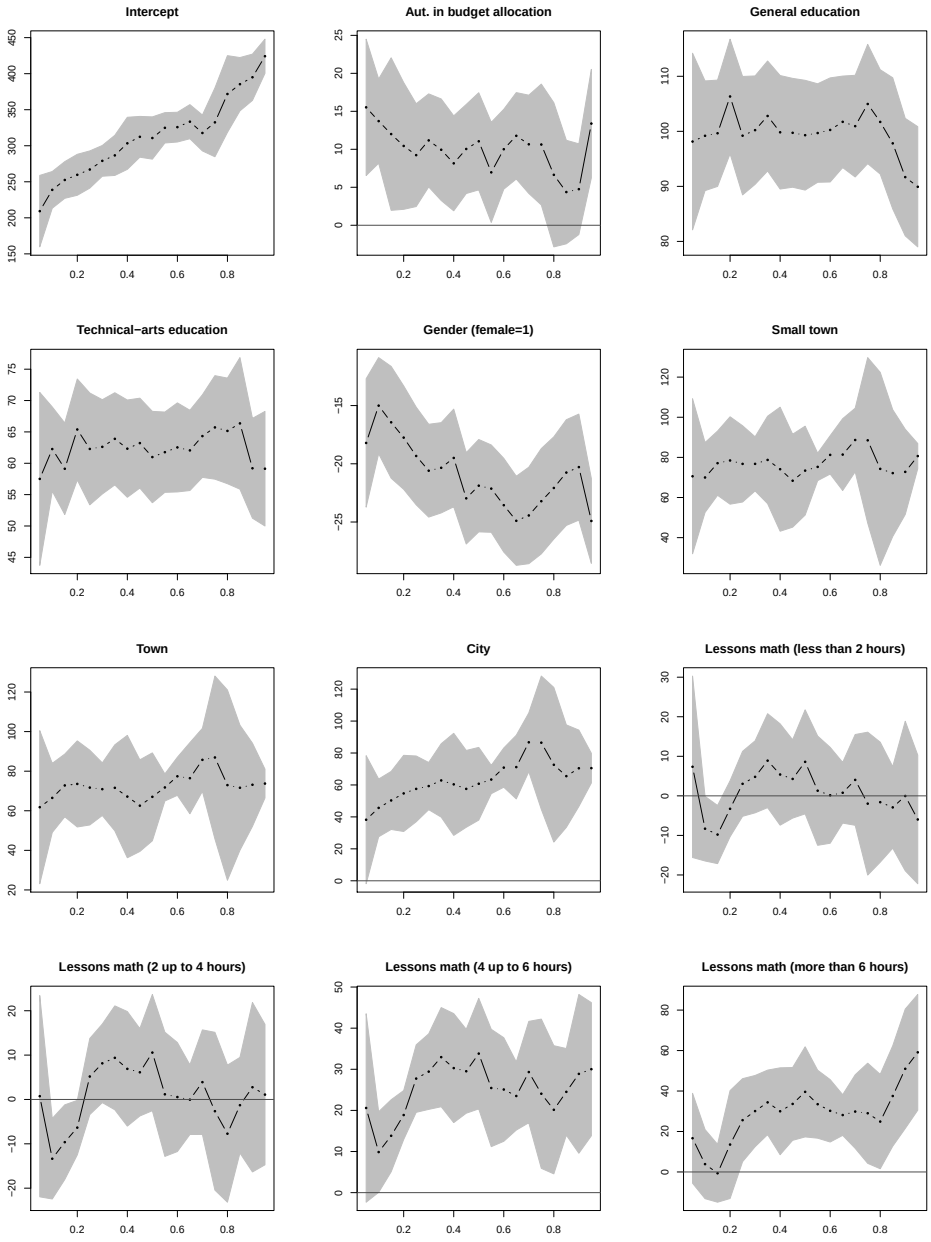


Figure 3.2: Quantile regression results - Part I

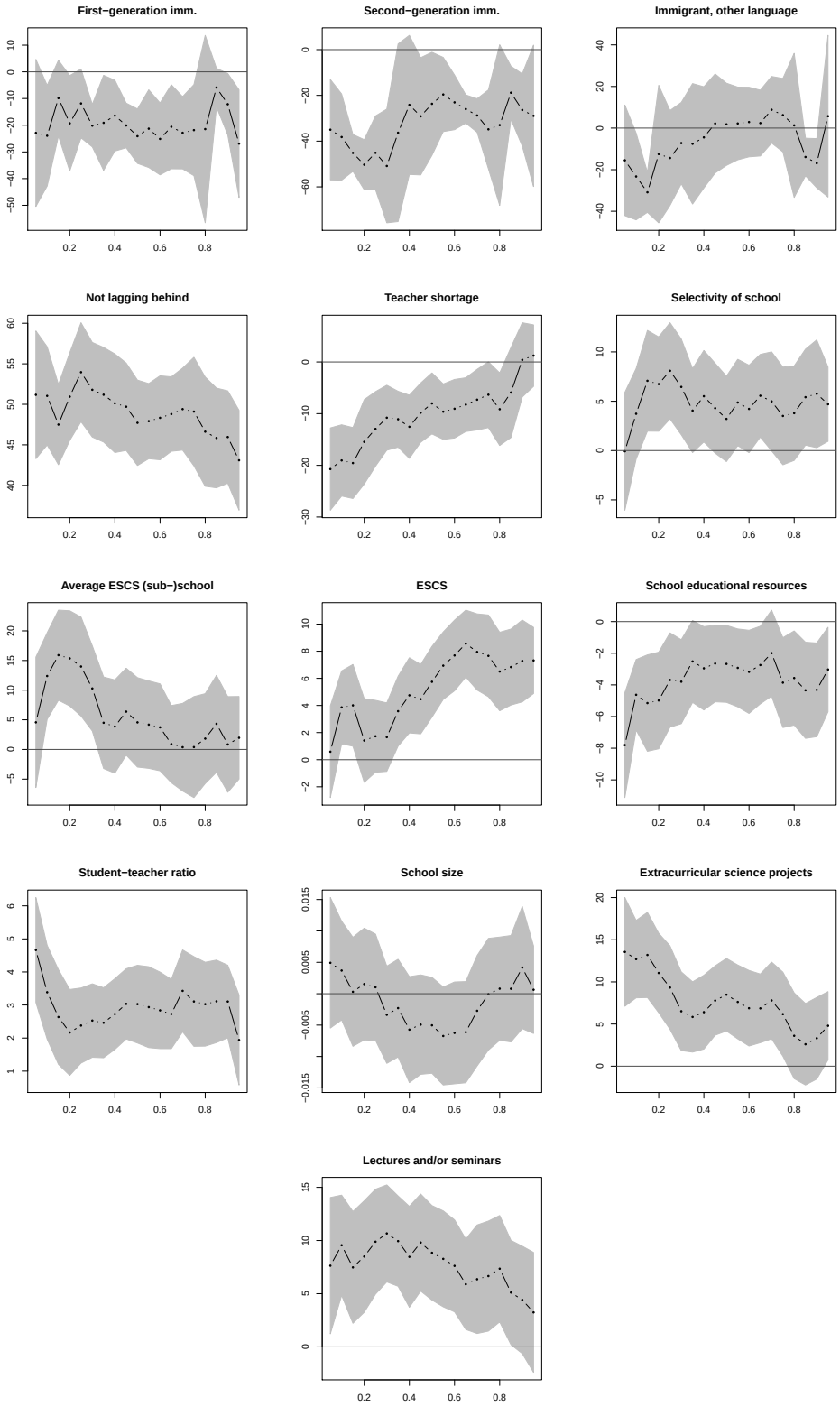


Figure 3.3: Quantile regression results - Part II

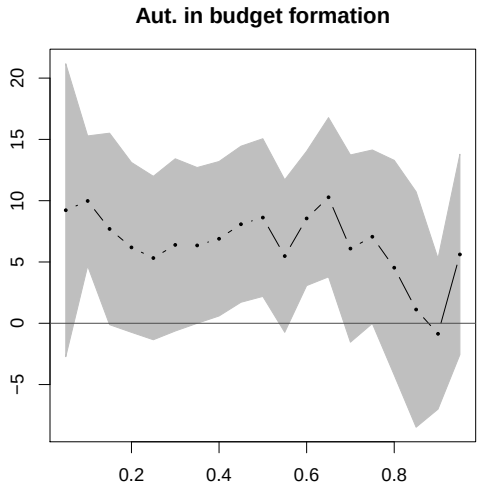


Figure 3.4: Quantile regression results of school staff autonomy in budget formation

References

Abdulkadiroglu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., 2011. Accountability and flexibility in public schools: evidence from Boston's charters and pilots. Quarterly Journal of Economics 126 (4), 2133–2134.

Ammermüller, A., 2004. PISA: What makes the difference? explaining the gap in PISA test scores between Finland and Germany. ZEW Discussion Papers 04-04, ZEW - Zentrum für Europäische Wirtschaftsforschung / Center for European Economic Research.

Besley, T., Case, A., 1995. Incumbent behaviour: vote-seeking, tax-setting, and yardstick competition. American Economic Review 85 (2), 25–45.

Bollaerts, K., 2009. Statistical models in epidemiology and quantitative microbial risk assessment applied to salmonella in pork. Ph.D. thesis, University of Hasselt.

Bollaerts, K., Eilers, P. H. C., van Mechelen, I., 2006. Simple and multipleP-splines regression with shape constraints. British Journal of Mathematical & Statistical Psychology 59, 451–469.

Bottani, N., Favre, B., 2001. School autonomy and evaluation. Prospects - Quarterly Review of Comparative Education 31 (4), 166, issue 120.

Card, D., Dooley, M., Payne, A., 2010. School competition and efficiency with publicly funded catholic schools. Canadian Labour Market and Skills Researcher Netwerk, Working Paper 66, 61.

Chubb, J., Moe, T., 1990. Politics, markets, and America's schools. Brookings Institution paper, 318.

Clark, D., 2009. The performance and competitive effects of school autonomy. Journal of Political Economy 117 (4), 745–783.

De Meyer, I., Warlop, N., 2010. PISA - Leesvaardigheid van 15-jarigen in Vlaanderen. De eerste resultaten van PISA 2009.

Deci, E., Ryan, R. M., 1985. Intrinsic motivation and self-determination in human behavior. New York: Plenum Press.

Eilers, P., Marx, B., 1996. Flexible smoothing with B-splines and penalties. Statistical Science 11, 89–121.

Eurydice, 2007. School autonomy in Europe. Policies and Measures. Eurydice, iSBN 978-92-79-07522-3.

Eurydice, 2008. Levels of autonomy and responsabilities of teachers in Europe. Eurydice.

Fan, C., Lin, C., Treisman, D., 2009. Political decentralization and corruption: Evidence from around the world. Journal of Public Economics 93 (1-2), 14–34.

Frey, B., 1993. Does monitoring increase work effort - The rivalry with trust and loyalty. Economic Inquiry 31 (4), 663–670.

Frey, B., 1994. How intrinsic motivation is crowded out an in. Rationality and Society 6 (3), 334–352.

Frey, B. S., OberholzerGee, F., 1997. The cost of price incentives: An empirical analysis of motivation crowding-out. American Economic Review 87 (4), 746–755.

Gu, C., Wahba, G., 1991. Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. SIAM Journal on Scientific and Statistical Computing 12 (2), 383–398.

Hallinger, P., Bickman, L., Davis, K., 1996. School context, principal leadership, and student reading achievement. Elementary School Journal 96 (5), 527–549.

Hanushek, E., Link, S., Wößmann, L., 2011. Does school autonomy make sense everywhere? Panel Estimates from PISA. CESifo Working Paper, No. 3648.

Hanushek, E. A., 2003. The failure of input-based schooling policies. Economic Journal 113 (485), F64–F98.

Hanushek, E. A., Luque, J. A., 2003. Efficiency and equity in schools around the world. Economics of Education Review 22 (5), 481–502.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. London: Chapman and Hall.

Hindriks, J., Myles, G., 2006. Intermediate Public economics. MIT Press.

Hindriks, J., Verschelde, M., Rayp, G., Schoors, K., 2010. School tracking, social segregation and educational opportunity: Evidence from Belgium. CORE Discussion Paper 2010/81.

Hirtt, N., 2007. Pourquoi les performances PISA des élèves francophones et flamands sont-elles si différentes ? Appel pour une école démocratique, 27.
URL `http://www.skolo.org/IMG/pdf/PISA_F.pdf`

Hoxby, C., 1999. The productivity of schools and other local public goods producers. Journal of Public Economics 74 (1), 1–30.

Hoxby, C., 2003. The economics of school choice. University of Chicago Press.

Hoxby, C. M., 2000. Does competition among public schools benefit students and taxpayers? American Economic Review 90 (5), 1209–1238.

Kane, T., Rockoff, J., Staiger, D., 2006. What does certification tell us about teacher effectiveness? Evidence from New York City. National Bureau of Economic Research Working Paper (12155).

Koenker, R., 2011. quantreg: Quantile Regression. R package version 4.67. URL `http://CRAN.R-project.org/package=quantreg`

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

Marx, B., Eilers, P., 1998. Direct generalized additive modelling with penalized likelihood. Computational Statistics and Data Analysis.

Niskanen, W., 1971. Bureaucracy and Representative Government. Chicago: Aldine-Atherton.

Niskanen, W., 1991. The budget-maximizing bureaucrat: Appraisals and evidence. In: Blais, A., Dion, S. (Eds.), A reflection on Bureaucracy and Representative Government. pp. 13–31.

OECD, 2005. PISA 2003 Data Analysis Manual: SAS users. OECD Programme for International Student Assessment.

OECD, 2006. PISA 2006: Science Competencies for Tomorrow's World. OECD Programme for International Student Assessment.

OECD, 2009. PISA 2006 Technical report. OECD Programme for International Student Assessment.

Poesen-Vandeputte, M., Bollens, J., 2008. Studiekosten in het secundair onderwijs. Wat het aan ouders kost om schoolgaande kinderen te hebben. Leuven: Katholieke Universiteit Leuven. Hoger instituut voor de arbeid.

Raudenbush, S. W., Bryk, A. S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods. Thousand Oaks, CA: Sage.

Rivkin, S. G., Hanushek, E. A., Kain, J. F., 2005. Teachers, schools, and academic achievement. Econometrica 73 (2), 417–458.

Shleifer, A., 1985. A theory of yardstick competition. RAND Journal of Economics 16 (3), 319–327.

Tiebout, C., 1956. A pure theory of local public expenditures. Journal of Political Economy 64, 416–424.

Wahba, G., 1980. Spline bases, regularization, and generalized cross validation for solving approximation problems with large quantities of noisy data. In: Cheney, W. (Ed.), Approximation Theory III. Academic Press, New York, pp. 905–912.

Wahba, G., 1990. Spline models for observational data. In: CBMS-NSF Regional Conference Series in Applied Mathematics. Vol. 59. Philadelphia: Society of Industrial and Applied Mathematics.

Wößmann, L., 2003. Schooling resources, educational institutions and student performance: The international evidence. Oxford Bulletin of Economics and Statistics 65 (2), 117–170.

Wößmann, L., 2008. Efficiency and equity of European education and training policies. International Tax and Public Finance 15 (2), 199–230.

Wößmann, L., Lüdemann, E., Schütz, G., West, M., 2007. School accountability, autonomy, choice, and the level of student achievement: International evidence from PISA 2003. Unclassified OECD Education Working Paper 13, 1–85.

Wood, S., 2006. Generalized Additive Models: an introduction with R. Texts in Statistical Science. Chapman and Hall/CRC.

Wood, S., Augustin, N., 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecological modelling 157, 157–177.

Wu, M., 2005. The role of plausible values in large-scale surveys. Studies in Educational Evaluation 31, 114–128.

4

School tracking, social segregation and educational opportunity: evidence from Belgium[1]

4.1 Introduction

“Tracking" is the grouping of students by ability between classes with differentiated curriculum, a strategy common in middle and high schools. The tracks cover distinctly different curricula across subjects, and lead to different destinations upon graduation. Three tracks are common: (1) a high track, with college-preparatory or honors courses that prepare students for admission

[1]This chapter is the result of joint work with Jean Hindriks (UCL), Glenn Rayp and Koen Schoors. Another version of this chapter appeared as CORE Discussion Paper 2010/81 and Ghent University Working Paper WP 10/690.

to top colleges and universities; (2) a middle track that served as a catch-all for the group of students in the middle, and (3) a low track, consisting of vocational courses and a smattering of low-level academic offerings, serving mainly low functioning and indifferent students. After graduation, low track students frequently drop out, go to work, or get unemployed.

One of the main reasons that tracking has become unpopular has less to do with the outcomes the practice generates than with the types of students who tend to be assigned to the different tracks. A major concern is that tracking is used to segregate students on the basis of family background and race, as well as ability. In fact, the primary charges against tracking are (i) that it doesn't accomplish anything and (ii) that it unfairly creates unequal opportunities for academic achievement. This critique has fueled in the 90's with the very influential book by Anne Wheelock (1992), '*Crossing the Tracks. How Untracking Can Save America's Schools*'.

Tracking, to be sure, links a student's present and past track level. As illustrated in Figure 4.1, if past academic achievement is related to parental background, then tracking will link present track to family background. As a result, students may be placed in low tracks because of the socio-economic status of their family.[2] If we believe that teaching follows a hierarchical sequence, exposing students to increasingly difficult skills and complex knowledge, early tracking can lock in students with low socio-economic background in low tracks and induce progressive segregation. The consequence is unequal access to knowledge. This is getting worse, as evidence seems to suggest, if low tracks attracts less experienced teachers and hinders the motivation and aspiration of students with lower expectations; and if parents intervention into tracking decision is more common with highly educated parents pushing for high track placements. This is where unequal opportunities comes into the debate.

[2]PIRLS 2006 displays clearly strong correlation between early reading literacy (grade 4) and parental education in all countries (Mullis et al., 2007, chapter 3, exhibit 3.5, pp. 120-121).

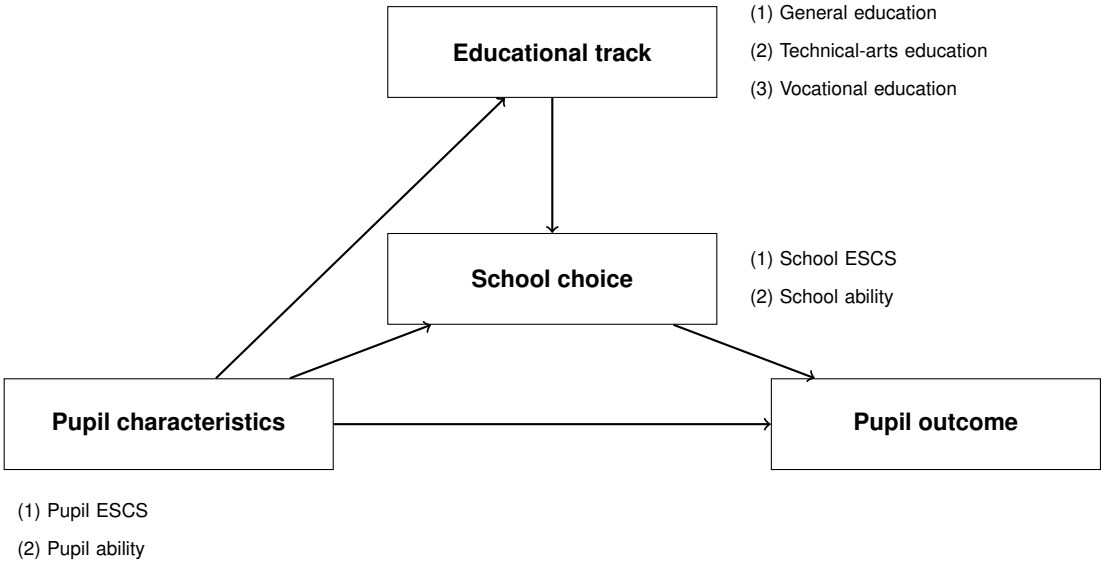


Figure 4.1: School tracking and inequality of opportunity

A first contribution of this paper is to estimate the relation between systematic school tracking and social segregation in schooling. To investigate empirically the link between school tracking and social segregation, we study systematic '*between-school*' tracking as is implemented extensively in among others Belgium, Hungary, Switzerland, Austria, Luxembourg, the Netherlands and Germany. Pupils are sorted in general (high), technical-arts (middle) and vocational (low – track) study programmes based on prior achievements, with almost no probability to have lessons together with pupils from different tracks and almost no possibility to share the same teachers. Each track has its own curriculum and end goals. Most schools organize only one track so that pupils in different tracks are in different schools. There is almost no upward track mobility.
This is clearly different from '*within-school*' tracking as discussed in Duflo et al. (2009) and is the case in among others the US[3]. We focus on the Flemish community of Belgium (Flanders), which has a long tradition of educational tracking at the age of 12 (grade 7). This is an interesting place to look at because it combines high average achievements (repeatedly in the top of international PISA Tests in Math, Science and Reading) with extensive achievements inequality between schools and between students. Belgium is also a country that displays very extensive so-

[3]Within-school tracking is discussed in among others Epple et al. (2002), Figlio and Page (2002).

cial segregation in education (see Jenkins et al. (2008)). Using Hutchens decomposition method, we show that most of social segregation in the Flemish community takes place across tracks. It is also shown that the private/public school dimension has little impact on social segregation.

A second contribution of this paper is to measure association between social segregation on educational opportunities in a cross-sectional microlevel study. We adopt recent empirical methods with strong theoretical underpinnings to study how school tracking relates to equality of opportunity in schooling. We estimate the existence of inequality of opportunity in schooling by comparing conditional (on socio-economic status) distributions of test scores and by estimating the 'Gini opportunity' index as in Lefranc et al. (2008). Then, the determinants of inequality of opportunity are investigated in a multilevel regression approach that is closely related to Bourguignon et al. (2007). This multilevel regression analysis relates the test score of students to the social composition of the school and the socio-economic status of students. To accommodate the hierarchical clustering of pupils in schools, we include school effects as well a specific random individual effect. We complement this analysis, with a broader conditional quantile regression analysis to see how not only the mean of the distribution, but also the full distribution is affected (possibly differently) by the family background and the school composition. Explicit considerations of these effects via quantile regression can provide a more subtle view of the stochastic relationship between socio-economic variables and test scores and therefore a more informative empirical analysis. Results show that social segregation has a strong association with inequality of opportunity in schooling. The quantile regression analysis reveals that the social composition of school is the main influence on the conditional quartiles of test scores. This means that students with the same family background achieve significantly different test score threshold depending with whom they go to school.

The remainder of the paper is structured as follows. In section 4.2, we review the related literature. In section 4.3, we present the data. In section 4.4, we discuss the methodology to estimate and explain inequality of opportunity, and to estimate and explain social segregation. In section 4.5, we present and discuss the results. In section 4.6 we provide some concluding remarks and discussions.

4.2 Literature on school tracking

Recent empirical research has shown the importance of educational achievements for (1) individual earnings, (2) the distribution of income, and (3) economic growth (Barro, 2001; Bishop, 1992; Nickell, 2004; Hanushek and Wößmann, 2008). As a result, the issue of equal earnings opportunities is closely related to equal educational opportunities (Brunello and Checchi, 2007). Understanding the drivers of equal opportunities in education is thus a major issue. Since the Coleman (1966) report, the impact of family background and peer effects on the quality of education has been investigated in a wide range of literature. Using data from respectively the US and Brazil, Betts and Roemer (2005) and Waltenberg and Vandenberghe (2007) show that not the redistribution of public budgets, but institutional features are the key in increasing equality of opportunity in schooling (see also the survey of Betts (2011)). As shown in OECD (2006), school composition is important in explaining educational achievement. In this paper, we study the causes and effects of social segregation in a within-country approach.

The study of school tracking is closely related to the study of peer effects[4]. In equilibrium analyses of among others Epple and Romano (1998), de Bartolome (1990), Benabou (1996), Nechyba (2000) and Epple et al. (2002), peer group effects are incorporated to study the impact of school vouchers, private-public school sorting and community structure. In empirical studies however, no consensus is reached on the relation between peer effects and educational outcomes (Brunello and Checchi, 2007). This, amongst others, because it is difficult to separate peer effects from other confounding effects. To overcome this problem, Hanushek et al. (2003) control for family background, school settings, student and school-by-grade fixed effects in a longitudinal panel study on a set of pupils in Texas. Hanushek et al. (2003) find evidence for a linear relationship between peer group quality and educational outcomes. Consequently, at an aggregate level, altering peer-group quality by educational tracking is expected to have no impact on the average education quality, but strong impact on inequality of educational outcomes. However, exploiting variation in test scores for the same pupil across subjects, Lavy et al. (2009) find evidence that

[4]The effect of other pupils that follow courses together with the student in question (i.e., the peers) on the educational achievement of the student.

only the top 5% and bottom 5% students matter to explain individual variations across subjects in secondary education in England. In addition, evidence is found that only academic achievement matters, not family background and that peer effects are heterogeneous in gender and student's ability.

In a recent randomized experiment in 121 primary schools in Kenya, Duflo et al. (2009) found that students in tracking schools scored 0.14 standard deviation higher on average than non-track schools. The random nature of the experiment is assured by the random attribution of teachers and schools in the two installed systems (tracking versus non-tracking) and by assigning pupils in two tracks by initial achievement (bottom halve of class selected into lower track, upper half to higher track). As both teachers and schools are randomly selected, there are no institutional effects by assumption. The effect of school tracking is found to be positive for both pupils in the low and high tracks. Obviously this experimental result need not extend easily to a more general (non-experimental) context. In fact, Guyon et al. (2010) find evidence against the efficiency argument of tracking. In a natural experiment, Guyon et al. (2010) investigate the impact of the 1989 'de-tracking' reform in secondary education in Northern Ireland with an increase of the relative size of pupils in elite track schools from around 31% to 35%. Information is used from 22 self-constructed areas in Northern Ireland for cohorts between 1974 and 1982. Discontinuity in educational outcomes is found across cohorts, showing a positive net-effect of de-tracking. In addition, a positive association between the area-level change in size of the reform and the area-level educational performance change shows the robustness of this positive net-effect of de-tracking.

There is also a vast literature of cross-country and country case studies showing the detrimental effects of school tracking for equality of opportunity in education. In Schutz et al. (2008), data from TIMMS and TIMMS-REPEAT is used to estimate and explain the effect of family background on student performance, using cross-country variations. Evidence is found that late tracking and a long pre-school cycle go hand in hand with lower impact of family background on educational attainment. Ammermüller (2005) and Hanushek and Wößmann (2006) use PISA and PIRLS data to estimate the effect of institutions on educational opportunities by applying a difference-in-difference approach. While Ammermüller (2005) deals with the effect of family

background, Hanushek and Wößmann (2006) discusses the effect of tracking on the distribution of test scores (i.e., inequality in outcomes). Both studies find evidence for a negative effect of educational tracking. Ammermüller (2005) finds in addition a negative effect of school systems with a large private school sector. However, using a difference-in-difference approach with data from PISA, TIMMS and PIRLS, Waldinger (2007) shows that the evidence from these cross-country difference-in-difference studies on educational tracking is vulnerable to specification changes. Using cohort data from 4 datasets, Brunello and Checchi (2007) find significant interaction between educational tracking and the effect of family background on educational attainment and early wages.

In the majority of country case studies - referenced in Brunello and Checchi (2007) - educational reforms that reduce educational tracking are found to be associated with less impact of family background on educational attainment. A well-known example is the case of Finland - where in 1972-1977 a two-track system was progressively replaced with a comprehensive school system till the age of 16. Pekkarinen et al. (2009) find in a difference-in-difference approach an increase of intergenerational income mobility as consequence of the reform. See also, for a similar result, the Swedish study of Meghir and Palme (2005).

The role of social segregation in explaining the effect of educational tracking on EOp in schooling is clarified by Checchi and Flabbi (2007). They show in a theoretical model of school-track choice that ability tracking can result in inequality of opportunity in schooling when the family background determines the track choices, given the level of ability (Contini et al., 2008).

Using the PISA 2000 and PISA 2003 dataset, Jenkins et al. (2008) find an association between social segregation and school choice for a sample of OECD countries. Using pupil and school-level data from the PISA 2006 dataset, Alegre and Ferrer (2009) find effect of social composition of schools on educational performance. They also find association between country-level social segregation and selection of pupils by schools and into ability tracks.

4.3 Data

The PISA 2006 dataset is used. In Belgium, education is organized by the Flemish community, the French-speaking community and the German-speaking community. We focus on the Flemish community. The PISA dataset is characterized by richness on variables related to educational achievement, family background and school level institutional settings. Although the main focus of PISA 2006 is on science, each participating pupil is asked to complete a standardized test on math, science and reading and fill out a survey with questions related to his/her family background, views on issues related to science, the environment, careers, learning time and teaching and learning approaches of science. Each principal of the participating schools is asked to complete a survey with questions on the characteristics of the school.

Tests are typically constructed to have assessed between 4500 and 10000 students of age 15 in each country. To sample the target population of 15-year old pupils that are at least in grade 7, PISA 2006 has implemented a two-stage stratified sample design. In a first stage, for each stratum[5], schools are sampled proportional to size from a list of schools in the region (PPS sampling). The target was 150 schools in each region. In a second-stage, 35 pupils are randomly drawn with equal probability from a list of 15-year old pupils in the school.[6] Final student weights are constructed to correct for varying selection probabilities of the students.[7] To incorporate sampling variation, a Balanced Repeated Replication (BRR) procedure with 80 replication estimates - described in OECD (2005)- can be used to construct standard errors (OECD, 2009). Alternatively, to do statistical inference, bootstrap resampling approaches can be used[8]. Both

[5]A group of schools, formed to improve the precision of sample based estimates.

[6]If the school size is lower than 35, all pupils are included in the sample.

[7]This occurs because of certain subgroups that are over- or under-sampled, the information of school size at the time is not completely correct, school non-response, student non-response and the inclusion of trimming weight to ensure stable estimates.(OECD, 2009)

[8]In this approach, for a sample of n observations, each bootstrap sample is a random sample of n observations selected with replacement from the original sample of observations. Consequently, in a bootstrap sample that is drawn with replacement, some of the n original sample observations are more than once in the bootstrap sample. Other original sample observations are not in an individual bootstrap sample. By replicating the bootstrap samples

approaches incorporate the final student weights.
PISA 2006 makes use of the plausible value approach to estimate the pupil performance in respectively mathematics, science and reading literacy. These plausible values are random values from the posterior distribution and may not be aggregated at pupil level (OECD, 2005). Therefore, in what follows, we use the first plausible value to estimate educational outcomes in math, science and reading at pupil level.

Pupils that are in special education or part-time education are deleted from the sample. By this, the sample is reduced to 4125 observations in the Flemish community of Belgium. Sub-schools are defined to investigate the importance of school tracking. A sub-school is defined as a unit that provides either general, technical-arts, or vocational education. When a school provides both general and technical or arts education (which is relatively rare), then the school is treated as two separate (sub-)schools.[9] The sample consists of 269 Flemish (sub-)schools.

Table 4.1 shows descriptive statistics of key variables. Standardized test scores for math, science and reading are high in the Flemish community (PISA average is 500). In addition, the high standard deviation of educational outcomes in the Flemish community shows that there is high inequality in individual test scores. To relate inequality in outcomes to family background, we consider 2 circumstance variables: socio-economic status and migration status. First, family socio-economic status is estimated by PISA as a composite index of the Economic and Socio-Cultural Status (ESCS) of a pupil, derived from (1) the highest occupational status of each student's parents, (2) their highest educational level, and (3) a summary measure of household possessions (OECD, 2009). Second, for migration status, three proxies are used. First-generation immigrants and second-generation immigrants are respectively defined as pupils that are not born in Belgium and pupils that are born in Belgium, but are children of immigrants. Pupils that are first- or second-generation immigrant and do not speak the school language at home are grouped in a latter category of non-native pupils. Table 4.1 shows that there are less than 8 per-

many times, we approximate the true population.

[9]This is necessary to construct 'between school track' and 'within school track' social segregation estimates later on

cent immigrant pupils in the sample. As mentioned above, only 35 pupils are sampled in each school. Consequently, a detailed look at segregation between migrants (small minority) and non-migrants (large majority) would lead to inaccurate estimates. Therefore, we do not study ethnic segregation in this paper. The focus is thus on social segregation and equality of opportunity.

In Belgium, secondary education starts in general at age 12 and ends at age 18. In the Flemish community, pupils in ordinary education are tracked in the first year of secondary education in general education, technical education, arts education and vocational education based on prior achievements. In our final sample, around 50 percent of pupils are in general education (high track), 28 percent are in technical-arts education (middle track) and 20 percent in vocational education (low track).[10]

If a pupil has not reached the basic skills, determined by the 'end goals' in a school year, grade repetition and re-orientation to lower tracks are used. In our final sample, 77 percent of pupils are 'on time'.

Private-granted schools are only a negligible proportion of the school population. There are mainly public schools (under control of community, provinces, cities or municipalities) and private operating, public-granted schools (e.g. Catholic schools, non-confessional schools). In our final sample, 74 percent of pupils are in private-operating schools.

[10]We merge the technical and arts tracks together because the two tracks do not dominate each other in curriculum difficulty and test scores and because there is only a small proportion of pupils that are in the arts track.

Variable	Mean	S.E.
Output		
PISA 2006 Performance in math, final sample (FS)	555.940	(3.054)
PISA 2006 Performance in reading, FS	537.757	(5.644)
PISA 2006 Performance in science, FS	541.023	(2.611)
PISA 2006 Standard deviation of performance in math, FS	89.190	(1.464)
PISA 2006 Standard deviation of performance in reading, FS	92.401	(1.793)
PISA 2006 Standard deviation of performance in science, FS	84.007	(1.372)
Circumstances		
Economic and Socio-Cultural Status (ESCS)	0.272	(0.025)
Proportion of first-generation immigrants	0.030	(0.006)
Proportion of second-generation immigrants	0.026	(0.005)
Imm. that speak non-off. Belgian language at home	0.022	(0.004)
Educational system		
School type (public=1, private-operating=0)	0.263	(0.019)
General education	0.479	(0.016)
Technical-arts education	0.325	(0.014)
Vocational education	0.196	(0.012)
Grade 10	0.771	(0.008)
Age of school tracking	12	
Number of observations	4125	

Notes: A SAS procedure for a Balanced Repeated Replication procedure with 80 replication estimates, described in OECD (2005), is used to construct the mean and standard error. Standard errors between brackets.

Table 4.1: Summary statistics

To provide descriptive information on the distribution of pupils from different family backgrounds across different tracks, we rank pupils in the sample by their ESCS level, and the top and bottom halves are assigned to the low and high social position groups.

Descriptive statistics in Table 4.2 indicate that pupils with low social position have significantly

less probability to attend general education (about two times less likely to attend the high track) and much higher probability to lag behind than pupils with a high social position (about twice more likely). Only 27 percent of pupils with a low social position are in general education without lagging behind. For pupils with a high social position, this is 58 percent. In addition, descriptive statistics indicate greater representation of pupils with low social position in public schools. In sum, pupils with a low social position have relatively higher probability to lag behind, to be in a lower track or to go in a public school.

Group	Pupils with low social position	Pupils with high social position
General education without lagging behind	0.272	0.575
	(0.251 , 0.293)	(0.552 , 0.594)
General education	0.313	0.643
	(0.293 , 0.334)	(0.621 , 0.664)
Technical or arts education	0.392	0.273
	(0.356 , 0.397)	(0.256 , 0.294)
Vocational education	0.310	0.085
	(0.289 , 0.330)	(0.071 , 0.096)
Lagging behind	0.283	0.160
	(0.262 , 0.304)	(0.143 , 0.177)
Public school	0.321	0.207
	(0.298 , 0.341)	(0.190 , 0.225)

Notes: Bootstrap approach with 999 replications and 95% basic confidence intervals between brackets, package 'boot' in R.

Table 4.2: Descriptive statistics on social segregation

4.4 Methodology

4.4.1 Defining inequality of opportunity

The main focus of the literature on equality of opportunity is on separating sources of inequality of outcomes that are morally acceptable and morally unacceptable.[11] It is argued that not all

[11]Seminal works are among others Arneson (1989), Barry (1991), Cohen (1989), Dworkin (1981a), Dworkin (1981b), Roemer (1993), Roemer (1998) and Sen (1980). For a discussion of differences and similarities between

inequality of outcomes is ethically immoral. Only inequality that is outside the realm of individual choice - referred as *circumstances* - should be eliminated by public intervention. On the other hand, inequality in outcomes that is a consequence of factors where individuals are judged to be responsible for are morally acceptable and should not be eliminated or compensated for.[12] Roemer (1998) assigns responsibility to individuals for the degree of *effort*. However responsibility can also be assigned to differences in preferences and individual choice.[13] There is a priori no reason to suspect that equality of outcomes and equality of opportunity goes hand in hand (Lefranc et al., 2008). Typical examples of circumstances are family background and individual attributes such as race, gender and place of birth. Examples of effort are own education, annual working hours and migration (Ramos and Van de Gaer, 2009).

Equality of opportunity can be measured *ex ante* or *ex post*. Ex post EOp occurs when individuals with the same effort obtain the same outcome and is measured as inequality within responsibility classes. As effort is unobserved, we measure ex ante EOp, which occurs when the opportunity sets are the same for all individuals, regardless of the circumstances. [14] Differently put, ex ante EOp is achieved when no particular vector of circumstances is preferred to another vector of circumstances by all individuals (Lefranc et al., 2008). As shown in Van de gaer (1993), Roemer (1998) and Lefranc et al. (2008), ex ante EOp amounts in comparing distributions of outcomes, conditional on circumstances.

Consider a situation where individuals are allowed to choose their circumstances s - referred as *type* - before they know their level of effort. EOp prevails between circumstances s and s' if s is not preferred to s' by all individuals, and vice versa. In other words, agents cannot order the opportunity sets of s and s'. The opportunity set of an individual can be represented by the conditional distribution function $F(x|s)$. Under the (weak) assumption that preferences satisfy

a capability approach as proposed by Sen (1980) and the followed equality of opportunity approach, see Fleurbaey (2006).

[12] See also Fleurbaey (2008a) and Fleurbaey (2008b) for a detailed account of equality of opportunity.

[13] Fleurbaey (1995) goes further in defining responsibility by delegation, which implies that individuals can be held responsible for outcomes that are not a part of any direct social objective.

[14] See e.g. Fleurbaey and Peragine (2009) and Checchi et al. (2010) for a discussion on ex ante and ex post EOp.

the criteria of first order stochastic dominance (FSD) and second-order stochastic dominance (SSD), stochastic dominance tests can be used to rank conditional distribution functions.[15] A formal definition of first order stochastic dominance (FSD) and second order stochastic dominance (SSD) is given in respectively (4.1) and (4.2).

Suppose inequality of opportunity where circumstance s is preferred to circumstance s' by all individuals. Inequality of opportunity defined as first order stochastic dominance between s and s' means that the distribution of outcome x conditional on s is for all x below the distribution of x conditional on circumstance s'.

However, it can easily be shown that this is a very weak definition of EOp. Indeed, suppose a situation where the outcome distribution of type s always dominates the outcome distribution of type s', except at the top (possibly when they exert maximal effort). Under the definition of inequality of opportunity as first order stochastic dominance, EOp is not rejected in this setting. But it is unfair because type s' must exert maximal effort to get a chance to outperform type s.

Second order stochastic dominance provides extra restrictions. Under the definition of inequality of opportunity as second order stochastic dominance - under the assumption of a Von Neumann-Morgenstern utility function- EOp prevails when the expected value derived from distribution $F(y|s)$ is not greater than the one derived from $F(y|s')$. However, SSD implies that equalization within a circumstance group is desirable, which is not necessarily the case.[16]

$$s \succeq_{FSD} s' \text{ iff } F(x|s) \leq F(x|s'), \forall x \in \Re_+. \tag{4.1}$$

$$s \succeq_{SSD} s' \text{ iff } \int_0^x F(y|s)dy \leq \int_0^x F(y|s')dy, \forall x \in \Re_+. \tag{4.2}$$

4.4.2 Measuring inequality of opportunity

Econometric stochastic dominance techniques can be used to test FSD and SSD of conditional distributions. A large advantage of this approach - against among other looking at the marginal effect of circumstances on the conditional mean of educational outcomes such as in Schutz et al.

[15] See Levy (1998) for an overview of the stochastic dominance approach.

[16] See e.g. Van de gaer et al. (2011) for a discussion.

(2008)- is that empirical tests have a strong theoretical underpinning[17]. However, stochastic dominance tests do not provide a complete ordering of EOp. An alternative we use that provides a complete ordering of EOp is a Gini-type index: the Gini Opportunity index as proposed by Lefranc et al. (2008). This index is based on the equivalence between SSD and generalized Lorenz dominance as shown by Shorrocks (1983). The Gini opportunity (GO) index is defined in (4.3)

$$GO(x) = \frac{1}{\mu} \sum_{i=1}^{k} \sum_{j>i} p_i p_j (\mu_j (1 - G_j) - \mu_i (1 - G_i)), \tag{4.3}$$

with k types, μ the mean of the population, μ_k the mean of group k, p_k the population weight of group k and G the Gini coefficient. The GO index computes the sum of all pairwise differences of the opportunity sets of all types, where the opportunity sets are defined as twice the area under Generalized Lorenz curve, $\mu_s(1 - G_s)$ for type s (Ramos and Van de Gaer, 2009). The GO index is in the interval $[0, 1]$. A value of 0 indicates full EOp. Bootstrapping can be used to do statistical inference.[18] However, as the GO-index is based on SSD, it also implies that equalization within a circumstance group is desirable, which is a hard assumption in an educational setting.

4.4.3 Explaining inequality of opportunity

Conditional mean regression approach

To estimate the effects of school factors and family background on student achievement, a multi-level regression analysis is carried out where covariates are distributed at two levels: the students and schools. In an educational setting, unobserved school effects are expected from school-level disparities in e.g. the academic culture of school staff. As students are clustered in different schools, the assumption of independent noise is violated. It is thus necessary to include school effects into the empirical analysis to obtain unbiased estimates (Raudenbush and Bryk, 2002).
The main purpose here is to identify two distinct channels for the impact of family background on individual educational attainments: a direct effect through the parental socio-economic sta-

[17] References of papers that use this approach to study equality of opportunity are among others Checchi et al. (2010), Lefranc et al. (2008), Peragine and Serlenga (2007) and Pistolesi (2009).

[18] See Davidson (2009) for a comparison of bootstrap, jackknife and asymptotic inference of the Gini index.

tus, and an indirect effect through school choice. In addition we will assess how much of the educational variation across schools can be explained by the school's average social position, after controlling for the rest of the individual family background and school level variables. As a result we can examine how and to what extent the social composition of schools (and so the social segregation between schools) relates to student achievement inequalities across schools.
To separate the direct and the indirect school choice effect of parental circumstances on individual achievement, we follow a regression approach closely related to Bourguignon et al. (2007). We consider that the educational outcome of an individual (O) is defined as a function of circumstance variables (C), effort variables (E) and unobserved determinants or random noise.[19] Individual ability is an unobserved variable and we do not possess information on ability scores or prior academic achievements. To allow for hierarchically clustered noise, we define θ_j as the random effect of school j and ε_i as the pupil-level errors. We further relax the i.i.d. assumption of ε_i by allowing for clustering within strata and by the introduction of probability weights to correct for unequal selection probabilities as proposed in Pfeffermann et al. (1998).

$$O_{ij} = f(C_i, E(C_i)) + \theta_j + \varepsilon_i \text{ , with i=1,...,n and j=1,...,m.} \tag{4.4}$$

Circumstances are supposed to have a 'direct' effect on outcomes and an 'indirect' effect via 'effort' (Bourguignon et al., 2007). The school choice (or track choice) is the only observable 'effort' variable we will consider (since other effort variables are not observables). The school choice is related to circumstances variables. Inequality of opportunity measures the overall effect of circumstances variables on educational attainments both via the direct effect of circumstances on educational achievements and the indirect effect through the effect of circumstances on the effort variables (school choice). The central feature of this paper is to include this indirect effect of circumstances on school choice via social segregation. For this, we include the school's average social position (S). Consequently, if we define n pupils in m schools, we obtain a model as

[19] Random noise is a combination of the effect of unobserved variables, measurement error and luck. We follow Lefranc et al. (2009) in defining luck as *"random determinants that are seen as a fair source of inequality provided that they are even-handed"*. Hence, we do not categorize random noise under effort or circumstances.

defined in (4.5).

$$O_{ij} = f(C_i, S_j) + \theta_j + \varepsilon_i \text{ , with i=1,...,n, j=1,..,m.} \tag{4.5}$$

The non-observable effort variables are captured by the unobserved determinants both at school levels θ_j and individual level ε_i. We specify a linear econometric model with varying intercepts as

$$\begin{aligned} O_{ij} =& \alpha_j + \beta C_i + \varepsilon_i, \\ \alpha_j =& \alpha + bS_j + \theta_j, \end{aligned} \tag{4.6}$$

where C_i and S_j are covariates at respectively the student and school level, ε_i and θ_j are independent errors at each level, β is a vector of coefficients for the circumstances C and b is the coefficient for the school's average social position S. Substituting the group level equation into the individual level equation gives the reduced form which can be estimated by maximum likelihood estimation as:

$$O_{ij} = \alpha + \beta C_i + bS_j + \theta_j + \varepsilon_i \text{ , with i=1,...,n, j=1,..,m.} \tag{4.7}$$

It is worth noting that estimates can be biased because of standard omitted variable problem due to the non-observable ability variable, and possible correlation between S and θ. If ability is correlated to circumstance variables C, then the residual terms are not orthogonal to the regressors. Therefore, in (4.7), we make the implicit assumption that (1) the social position of a pupil is unrelated to his true " ability" (no genetic transmission of cognitive ability) and (2) that the school's average social position is independent of the random school effects. The later assumption rules out the possibility that rich parents are more likely to have their children taught by better teacher than poor students. We cannot include a fixed school effect in the model because the school composition is already a school-level variable. So we use the random school effect to control for unobserved school features that are random and not systematically related to social class.

Alternatively, one could think of using instrumental variable approach. However, this strategy is not promising in our educational setting, because it is unlikely to find a variable that is correlated

to circumstance variables and has no direct effect on educational outcome (Bourguignon et al., 2007). Alternatively, following Bourguignon et al. (2007), one could explore the magnitude of the potential biases by a monte-carlo approach where a wide range of correlations between the residual terms and covariates are explored. However, extension of the proposed approach from OLS to a two-stage maximum likelihood estimation procedure is not pursued in this paper. We postpone to the concluding section the discussion of the implication of the omitted ability variable on our analysis and results.

Conditional quantile regression approach

To obtain a more complete picture of the conditional distribution of pupil performance, Koenker and Bassett (1978) introduced the estimation of conditional quantiles rather than the conditional mean.[20] Following this method, the conditional αth quantile ($\alpha \in (0,1)$) is defined as the test score threshold such that α percent of the pupils of the reference group perform worse and $1-\alpha$ percent perform better. It is given by the inverse of the conditional CDF:

$$q_\alpha(x) = \inf\{y : F(y|x) \geq \alpha\} = F^{-1}(\alpha|x). \tag{4.8}$$

For example, if y is the pupil test score and x her socio-economic status, 25 % of the pupils in the same reference group x performs worse than the score threshold $q_{0.25}(x)$. Recently, Li and Racine (2008) proposed a nonparametric kernel approach to estimate conditional CDF functions in a multivariate setting with both continuous and discrete variables.[21] By using kernel weights, no a priori parametric assumptions need to be imposed on the quantile regression. The main features of the Li and Racine (2008) approach in comparison to other kernel quantile regression approaches are that 1) it admits smoothing of both discrete and continuous covariates, 2) irrelevant variables are 'smoothed out' with high probability via the data-driven bandwidth selection, 3) also the dependent variable can be smoothed to improve estimation and 4) optimal bandwidths

[20] See Koenker (2005) for an overview of parametric quantile regression approaches

[21] This approach is implemented in the programing software R as package 'np' of Hayfield and Racine (2008). See Li and Racine (2007) for an excellent overview of nonparametric econometrics.

are selected in an automatic data-driven approach.[22] A disadvantage of this nonparametric quantile regression approach is that it is - to our knowledge - not possible to include random school effects.[23] The smooth estimate of $F(y|x)$ that allows the inclusion of mixed discrete ($X^d \in \Re^r$) and continuous covariates ($X^c \in \Re^q$) can be defined as:

$$\hat{F}(y|x) = \frac{n^{-1}\sum_{i=1}^{n} G(\frac{y-Y_i}{h_0}) K_\gamma(X_i,x)}{\hat{\mu}(x)}, \tag{4.9}$$

where $\hat{\mu}(x) = n^{-1}\sum_{i=1}^{n} K_\gamma(X_i,x)$ is the kernel estimate of $\mu(x)$. h_0 is the smoothing parameter associated with y. With generalized product kernel $K_\gamma(X_i,x) = W_h(X_i^c,x^c)L_\lambda(X_i^d,x^d)$. Product kernel of X^c is $W_h(X_i^c,x^c) = \prod_{s=1}^{q} h_s^{-1} w((X_{is}^c - X_s^c)/h_s)$, with $w(\cdot)$ a univariate kernel function and h a smoothing parameter associated with X^c. Product kernel of X^d is $L_\lambda(X_i^d,x^d) = \prod_{s=1}^{r} l(X_{is}^d,x_s^d,\lambda_s)$, with univariate kernel function $l(X_{is}^d,x_s^d,\lambda_s) = \mathbf{1}(X_{is}^d = x_s^d) + \lambda_s \mathbf{1}(X_{is}^d \neq x_s^d)$ and smoothing parameter $\lambda \in (0,1)$. $G(\cdot)$ is a CDF, e.g. the standard normal CDF. To obtain the conditional quantile estimate $\hat{q}_\alpha(x)$, we minimize the following objective function:[24]

$$\hat{q}_\alpha(x) = \arg\min_q |\alpha - \hat{F}(q|x)|. \tag{4.10}$$

4.4.4 Social segregation

Social segregation - that is the uneven distribution of social groups across schools - can be represented by a segregation curve or as a numerical measure.[25] The main drawback of segregation curves is that only an incomplete or partial ordering is provided. If segregation curves intersect, the segregation curve approach is silent about which distribution is more segregated (Hutchens, 2004). Unlike segregation curves, cardinal measures produce complete rankings. It is shown in

[22] As recommended in Li and Racine (2008), we use least-squares cross-validation for bandwidth selection. A second-order Gaussian kernel is used for nonparametric weighting.

[23] Inclusion of fixed school effects is not possible in this setting because it would induce an identification problem between the school level covariate S and the school level fixed effects (analogously to the multicollinearity problem in parametric models).

[24] An alternative is to use a check function approach as described in Koenker (2005) and Li and Racine (2007).

[25] A segregation curve plots the cumulative fraction of type 1 people (vertical axis) and type 2 people (horizontal axis), both fractions being ranked from low to high values of x_{1j}/x_{2j}, with x_{1j} (and x_{2j}) respectively the fraction of type 1 (2) in school j (Hutchens, 2004).

Hutchens (2004) that a square root index is the only measure of segregation that satisfies seven properties needed for a complete and additively decomposable ordering.[26] We use the Hutchens (2004) square root index to study social segregation in education - this is the segregation between socio-economic groups in schooling.

The Hutchens (2004) square root index is defined as the sum, over all schools, of each school's gap from proportional representation. As formulated in (4.11), for each school, this gap is the difference between the geometric mean of the proportional representation of children from different reference groups and the geometric mean of the actual group proportions. (Jenkins et al., 2008)

$$H = \sum_{i=1}^{S} \left[\sqrt{\frac{p_i}{P} * \frac{r_i^{\text{no seg}}}{R}} - \sqrt{\frac{p_i}{P} * \frac{r_i}{R}} \right], \tag{4.11}$$

with p_i the number of children with a low social position in school $i = 1,...,S$. Low social position can be defined as the first, second or third quartile of ESCS distribution. r_i is the number of children with a high social position in school $i = 1,...,S$. P and R are the proportions of children in the population with respectively a low and a high social position. If the low social position is defined as the first quartile, then $P = 0.25$ and $R = 0.75$. If there is no segregation, there is proportional representation of each group in each school so that $\frac{p_i}{P} = \frac{r_i}{R}$ in every school. Thus, (4.11) can be written as (4.12).

$$H = \sum_{i=1}^{S} \left[\frac{p_i}{P} - \sqrt{\frac{p_i}{P} * \frac{r_i}{R}} \right]. \tag{4.12}$$

[26](1) *Scale invariance*: if N_1 and/or N_2 are multiplied by a positive scalar and the share of both types in the S schools remains the same then segregation does not change, with N_1 (N_2) the number of observations of type 1 (type 2) in the schools. (2) *Symmetry in groups*: measure is unaffected if all the people in group i trade places with those in group j. (3) *Movement between groups*: social segregation increases when there is a 'disequalizing' movement of a student between schools. For example, if a pupil with a low social position in a 'rich' school moves to a 'poor' school, segregation increases. (4) *Insensitivity to proportional divisions*: division of a group into subgroups through a proportional division should not alter segregation. (5) *Additive decomposability*: the segregation measure can be decomposed in as sum of between-group segregation and within-group segregation.(6) *Symmetry in types*: the segregation measure is unaffected if pupils with a low social position are named group 1 or group 2. (7) *Index in interval* $[0,1]$, with 0 no segregation and 1 full segregation. (Hutchens, 2004)

By the additive decomposability property, (4.12) can be written as (4.13). Social segregation can be decomposed as the sum of between-group segregation ($H_{between}$) and within-group segregation (H_{within}).

$$H = H_{within} + H_{between} \text{ ,where } H_{within} = \sum_{g=1}^{G} w_g H_g, \text{ with } w_g = (P_g/P)^{0.5}(R_g/R)^{0.5}, \tag{4.13}$$

with $g = 1,...,G$ groups, w_g the weight of group g, P_g and R_g the number of pupils in group g with respectively a low and high social position.

As in Jenkins et al. (2008), as robustness check, the results are compared to estimates with the Duncan and Duncan (1955) dissimilarity index, also called the 'displacement index'. In this setting, the Duncan and Duncan (1955) dissimilarity index measures the fraction of pupils with low social position that would need to be displaced to 'rich' schools, without replacing them by other children, in order that every school has the same proportions of children with low and high social background.(Duncan and Duncan, 1955)

$$D = \frac{1}{2}\sum_{i=1}^{S}\left|\frac{p_i}{P} - \frac{r_i}{R}\right|. \tag{4.14}$$

4.5 Empirical results

4.5.1 The extent of inequality of opportunity

In this section, we use a conditional distribution approach to investigate the existence of inequality of opportunity in schooling. We study inequality of opportunity in schooling, caused by socio-economic status (ESCS). Figure 4.2 shows that the distribution of pupils in higher ESCS quartiles dominates by FSD the distribution of pupils in lower ESCS quartiles. This indicates that there exists inequality of opportunity in schooling associated to family background.

Table 4.3 shows the results from the Gini Opportunity (GO) estimates. In this univariate analysis, inequality of opportunity between pupils in 4 quartiles of ESCS is studied. For this, the dependent variable is the first principal component (FPC) of pupil performance on the PISA

2006 standardized tests for math, science and reading. Higher values indicate higher inequality of opportunity in schooling.

In the Flemish community, a Gini Opportunity of 0.016 is found, which is significantly different from zero. Consequently, we find evidence for a significant inequality of opportunity in schooling, caused by ESCS.

The robustness of this result is shown by estimation of EOp between 2 equal quantiles groups and 6 equal quantiles groups.[27] In addition, to test the sensitivity of the results for inclusion of migration status, we estimate Gini Opportunity for a sample of native pupils.

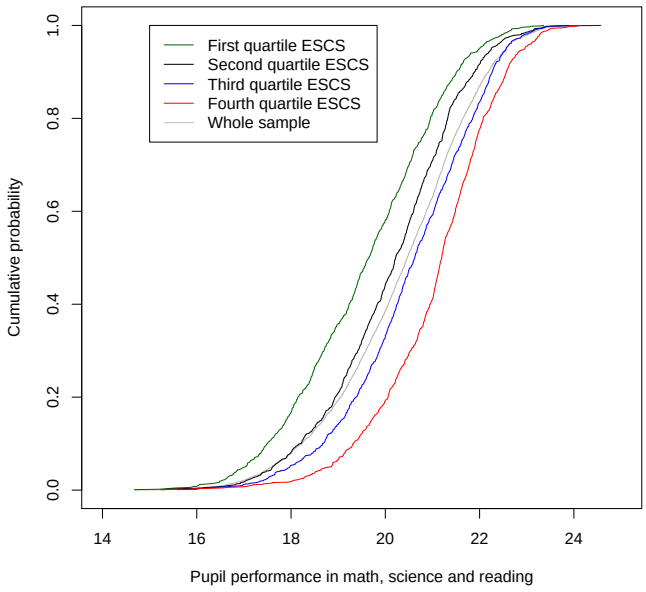


Figure 4.2: Conditional distribution of pupil achievement

[27] We thus split the groups at respectively the ESCS median and at ESCS quantiles (1/6 , 1/3, 1/2, 2/3, 5/6).

Variable	Estimates
Gini opportunity (GO)× 100 - 4 groups	1.647
	(1.522 , 1.832)
Gini opportunity (GO)× 100 - 2 groups	1.270
	(1.152 , 1.416)
Gini opportunity (GO)× 100 - 6 groups	1.693
	(1.545 , 1.843)
Gini opportunity (GO)× 100 - 4 groups - native pupils	1.560
	(1.413 , 1.709)

Notes: Bootstrapping with replacement, 999 replications, package 'boot' in R, 95 % basic confidence intervals between brackets.

Table 4.3: Gini opportunity estimates

4.5.2 Social segregation and inequality of opportunity

The conditional mean regression

Before discussing the indirect effect of circumstances as in Bourguignon et al. (2007), we show that a large proportion of between-school variation in outcomes is related to social segregation. OECD (2006) shows that in the Flemish community, 38.6 percent of the between-school variance in science performance and 1.9 percent of the within-school variance are explained by respectively the ESCS of schools and pupils. The OECD average is 20.5 and 3.8 percent respectively. So there are strong indications that the school ESCS has a large impact on inequality of opportunity in Belgium. This is in line with Jenkins et al. (2008), where Belgium is ranked as a high social segregation country. By implementing the Hutchens (2004) square root index on the PISA 2000 and 2003 datasets, Jenkins et al. (2008) show that only Hungary has higher social segregation than Belgium (H=0.142) in a sample of 30 OECD countries.

However, the effect of school ESCS is underestimated in OECD (2006). In the Flemish commu-

nity, different tracks can coexists. Because the peer effects and institutional effects are mainly influenced by the situation within a given track, we expect a larger effect if we study the effect of sub-schools ESCS. For this - as mentioned in section 4.3 - we treat separately different tracks in the same school as different sub-schools.

We extend the OECD (2006) results in two ways. First, we allow for random (sub-)school effects by estimating a two-level regression model as discussed in section 4.4. As in OECD (2006), we choose an additive linear functional form (HLM model). With level 1 the pupil and level 2 the sub-school. Second, instead of investigating pupil performance in science, we proxy educational outcomes by the standardized first principal component of math, reading and science.

Results in Table 4.4 show first, that only 2.87 percent of within-school variation in performance is explained by ESCS and the proxies for migration status. If only ESCS of the pupils is considered, this is only 0.43 percent. The significant coefficients and explanatory power of the proxies of migration status indicate that migration status matters. In the Flemish community, non-native pupils perform significantly worse than other native pupils.

Second, and more related to the issue of social segregation, the results show that school ESCS can explain around 60 percent of the between-school variation. Pupils in a school with many pupils with an unfavorable family background perform on average lower. The school's average ESCS should not be interpreted as a pure peer effect, but rather as the effect of social segregation, which includes both peer effects and institutional effects. The large explanatory power of school ESCS indicates that social segregation with large variation in school ESCS is a powerful predictor of variation in educational achievements between schools.

Regarding the indirect effect of circumstances. We find a large coefficient for school ESCS- respectively 1.70, which implies a 0.75 standard deviation disparity in educational outcomes between the first and third quartile school ESCS. Differently put, the large coefficient of school ESCS indicates large indirect effects of circumstances via the (self-)selection of pupils with a low (high) social position in schools where the social position of pupils is on average low (high). Obviously, it is not possible to say whether they perform better because of the positive influence in the classroom (endogenous effect) or because they share similar unobserved favorable

characteristics (correlated effect) (see Manski (1993)).[28]

Variable	Model I	Model II
ESCS of pupil	0.121***	0.102***
	(0.026)	(0.024)
Sub-school average ESCS	1.752***	1.701***
	(0.117)	(0.121)
First-generation immigrant		-0.429***
		(0.126)
Second-generation immigrant		-0.443***
		(0.132)
Immigrant that does not speak official		
Belgian language at home		-0.568***
		(0.133)
Log likelihood	-5966.299	-5919.216
Between-sub-school variation explained	59.532%	59.100%
Within-sub-school variation explained	0.428 %	2.869%
Number of level 1 units	4125	4125
Number of level 2 units	269	269

Significance levels : $*$: 5% $**$: 1% $***$: 0.1%

Notes: Dependent variable: first principal component of first plausible value of test scores reading, math and science. Standard errors between brackets. The model, estimated in STATA with GLLAMM, is a two-level model with level 1 the pupils and level 2 the sub-schools. We allow for clustering within strata. Final student weights are introduced as probability weights as proposed in Pfeffermann et al. (1998). To obtain between- and within-sub-school variation explained, we compare the model with only a constant with the model with explanatory variables, as suggested in OECD (2009).

Table 4.4: Multilevel regression

[28]Gender and interaction between individual ESCS and school ESCS were excluded because no significant effect was found. Results on the effect of social segregation are robust for altering the definition of school composition to first-quartile, median and third quartile school ESCS and for the introduction of social diversification in the model.

The conditional quantile regression

The nonparametric quantile regression approach brings finer results. In a preliminary regression, we constructed a model with 1 dependent variable (FPC of math, science and reading), 3 continuous covariates (school ESCS, pupil ESCS and social diversification index) and 3 discrete variables (dummy for first-generation immigrant, second generation immigrant and immigrant that does not speak official Belgian language at home). The dummies for first-generation and second generation immigrants were 'smoothed out' and are thus estimated to be irrelevant. In addition, no (interaction) effects was found for within-school diversification. Therefore, we re-estimated the nonparametric model with only 2 continuous covariates (ESCS and school ESCS) and 1 discrete covariate (Language at home). Figure 4.3 shows that we only find a small positive effect of pupil ESCS on the conditional first quartile test score($q_{0.25}$), median test score($q_{0.50}$) and third quartile test score ($q_{0.75}$). For the school ESCS, profound effects on pupil performance are found. Figure 4.4 shows a strong positive association between the three respective conditional quantiles and the school ESCS. The finding of a strong effect of school ESCS is thus robust for altering the estimation methodology. Figure 4.6 illustrates further that the effect of ESCS is almost completely captured by the indirect effect of social segregation (school ESCS). Only where school ESCS is low and individual ESCS is high - and the data is thus very sparse- we find a positive effect of pupil ESCS. For the language spoken at home, we find small, but negative effects on the conditional quantiles (see Figure 4.5).

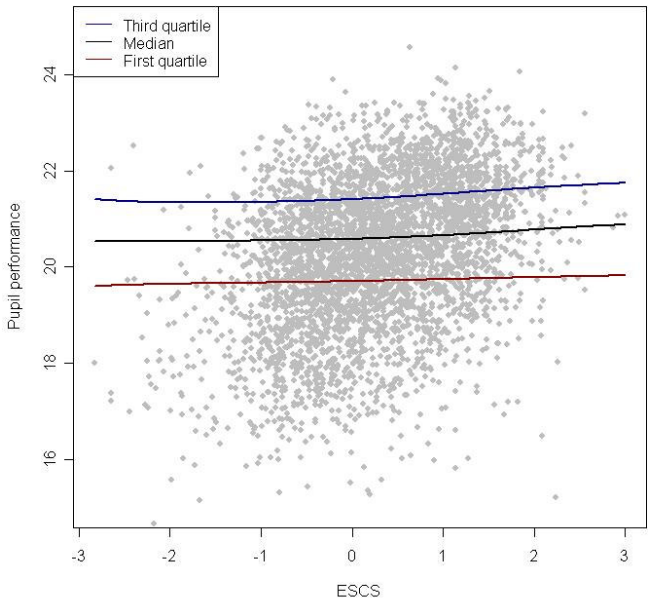


Figure 4.3: Condtional quantile estimates: effect ESCS

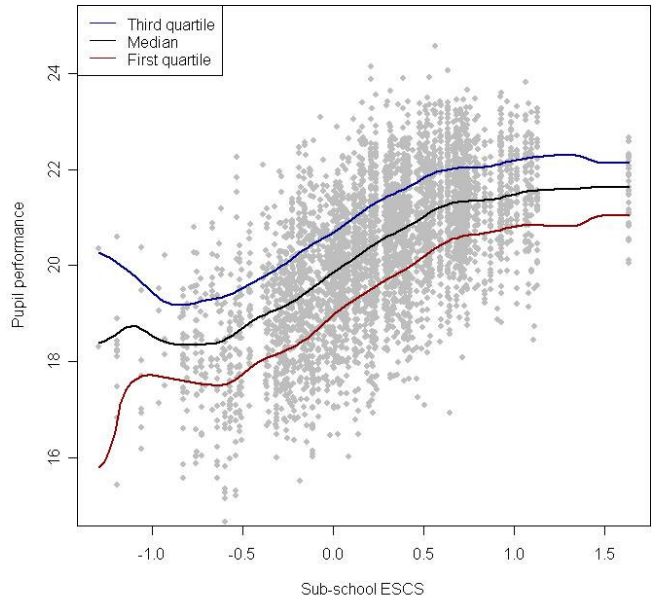


Figure 4.4: Conditional quantile estimates: effect sub-school ESCS

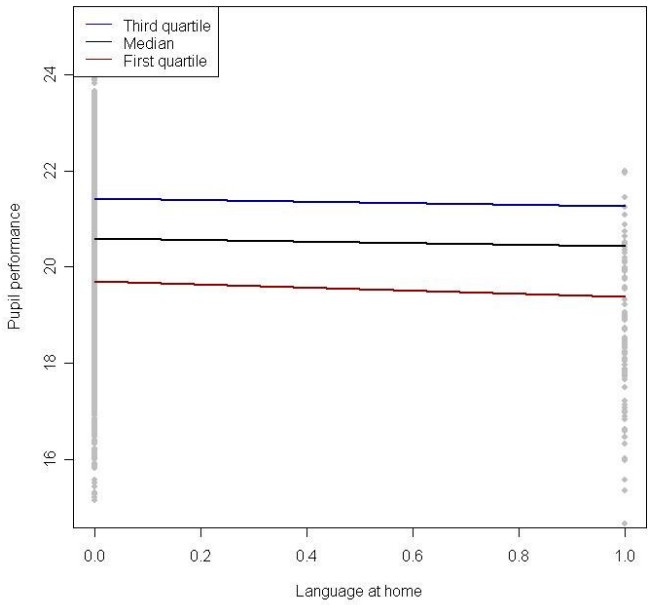


Figure 4.5: Conditional quantile estimates: effect language at home

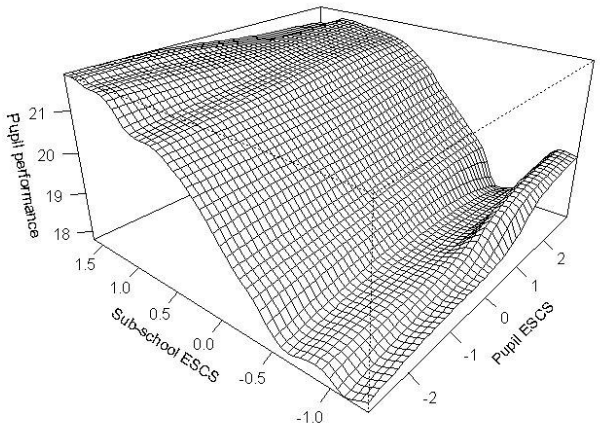


Figure 4.6: Conditional quantile surface: effect ESCS and school ESCS on median output

4.5.3 The impact of tracking on social segregation

Pupils have a low social position if they have ESCS below the median and they are with a high social position otherwise. In the descriptive statistics, we have shown that pupils with a low social position are unevenly distributed (1) between public and private-operating publicly funded schools and (2) across school tracks (general, technical-arts and vocational). In this section, we decompose social segregation into school types and into school tracks. For this, we decompose the Hutchens (2004) square root index as in (15) and (4.16). This with w the weight of the given subgroup, H_{group} the within-group Hutchens (2004) square root index of social segregation and $H_{between}$ the between-group social segregation.

$$H = w_{general}H_{general} + w_{technical-arts}H_{technical-arts} + w_{vocational}H_{vocational} + H_{between} \quad (4.15)$$

$$H = w_{public}H_{public} + w_{private}H_{private} + H_{between}. \quad (4.16)$$

Table 4.5 shows that social segregation prevails. A Hutchens (2004) square root index (H) is found of 0.14 and a Dissimilarity index (D) of 0.39. This latter index is easy to interpret: 39 percent of pupils with low social position need to be displaced from 'poor' to 'rich' schools-without replacing them by other children - in order that every school has the same share of children with low and high social background.

School tracking is found to be a main driver of social segregation. Indeed, between-track segregation explains 50.89 percent percent of social segregation in schools. Decomposition of the Hutchens (2004) square root index shows that only 6.56 percent of social segregation can be explained by school type.

In sum, we find evidence by decomposition that social segregation is for a large part driven by school tracking.

As robustness test, we define differently pupils into low social positions as those in the first quartile ESCS. Results are given in Table 4.6. We find similar results with some variation in the size of the tracking impact on social segregation. The main conclusion that school tracking is

far more important than school type in explaining social segregation is robust. However, social segregation between school types is now significantly different from zero.

Segregation index	Estimates
Square root index (H)	0.135
	(0.115 , 0.158)
Dissimilarity index (D)	0.389
	(0.358 , 0.425)
In general education (H_{general})	0.060
	(0.044 , 0.077)
In technical-arts education ($H_{\text{technical-arts}}$)	0.051
	(0.038 , 0.064)
In vocational education ($H_{\text{vocational}}$)	0.144
	(0.098 , 0.192)
Within track segregation (H_{within})	0.066
	(0.055 , 0.080)
Between track segregation (H_{between})	0.069
	(0.051 , 0.085)
Within track segregation (H_{within} as % of H)	49.1 %
	(42.2 , 56.5)
Between-track segregation(H_{between} as % of H)	50.887 %
	(43.5 , 57.8)
In public schools (H_{public})	0.147
	(0.106 , 0.192)
In private-operating schools (H_{private})	0.121
	(0.097 , 0.146)
Within school type segregation (H_{within})	0.127
	(0.107 , 0.148)
Between school type segregation (H_{between})	0.009
	(-0.001 , 0.014)
Within school type segregation (H_{within} as % of H)	93.4 %
	(89.4 , 100.5)
Between school type segregation(H_{between} as % of H)	6.6 %
	(-0.5 , 10.6)
Sample of sub-schools	269
Sample of pupils	4125

Notes: Bootstrapping with replacement, 999 replications, package 'boot' in R. 95% basic confidence intervals between brackets.

Table 4.5: Decomposition of social segregation

Segregation index	Estimates
Square root index (H)	0.168
	(0.139 , 0.200)
Dissimilarity index (D)	0.413
	(0.377 , 0.453)
In general education (H_{general})	0.159
	(0.111 , 0.210)
In technical-arts education ($H_{\text{technical-arts}}$)	0.075
	(0.053 , 0.095)
In vocational education ($H_{\text{vocational}}$)	0.105
	(0.065 , 0.143)
Within track segregation (H_{within})	0.109
	(0.090 , 0.132)
Between track segregation (H_{between})	0.059
	(0.039 , 0.075)
Within track segregation (H_{within} as % of H)	65.0 %
	(57.8 , 73.8)
Between-track segregation(H_{between} as % of H)	35.0 %
	(26.2 , 42.2)
In public schools (H_{public})	0.145
	(0.086 , 0.195)
In private-operating schools (H_{private})	0.160
	(0.128 , 0.197)
Within school type segregation (H_{within})	0.154
	(0.124 , 0.181)
Between school type segregation (H_{between})	0.014
	(0.003 , 0.024)
Within school type segregation (H_{within} as % of H)	91.4 %
	(86.0 , 97.9)
Between school type segregation(H_{between} as % of H)	8.6 %
	(2.1 , 14.0)
Sample of sub-schools	269
Sample of pupils	4125

Notes: Pupils with low social position are defined as pupils with an ESCS score below the first quartile of ESCS in the region. Bootstrapping with replacement, 999 replications, package 'boot' in R. 95% basic confidence intervals between brackets.

Table 4.6: Decomposition of social segregation - sensitivity analysis

4.6 Concluding remarks and discussions

In this paper we have investigated the importance of school tracking for inequality of opportunity in education. For this, we used Flemish pupil and school level data from the PISA 2006 dataset. First, we have shown the existence of inequality of opportunity in schooling by stochastic dominance testing on conditional distributions and using a bootstrapped version of the Gini Opportunity index. Second, we showed in a two-level regression (with school specific effect) that social segregation is a main driver of inequality of opportunity in schooling in Flanders. Over 60 percent of the variation between schools in educational outcomes can be explained by the variation in the school social composition. Using a conditional quantile regression approach, we showed that conditional quantiles are mostly influenced by the school socio-economic composition, with almost no influence of individual socio-economic status. This result suggests that the individual school opportunity set depends much more on the school memberships than on the individual family background. Lastly, to link tracking to social segregation, we decomposed the Hutchens square root index of social segregation between tracks and within tracks. We find strong evidence for a crucial role of school tracking in explaining social segregation. Only 6.56 percent of social segregation can be explained by school type while the between-track segregation explains 50.89 percent of social segregation.

Although this paper shows the high association between social segregation, inequality of opportunity and school tracking, we cannot really provide a causal relation. The fundamental reason is that we cannot control for unobserved ability levels. For instance, it could be that there is no social bias in the track assignment after controlling for the cognitive ability of the students. Moreover the regression estimates can overestimate the impact of family background on (average) test scores if (unobserved) cognitive ability is positively correlated to parental educations. In fact it is now well documented that there is parental transmission of cognitive ability. See Holmlund et al. (2008), Plug and Vijverberg (2003) and various articles in Nature magazine. Our response to this genetic transmission issue is threefold. First, it is fair to say, that there is a risk that some would consider social inequality in education achievements as natural because of the genetic transmission of ability. Second, we have been less preoccupied with the ability trans-

mission in this work because we have concentrated mostly on explaining average differences between groups and not between individuals. We have shown that the social composition of the school is a very powerful predictor of individual test scores. To keep this result in perspective, it should be emphasized that average differences in cognitive ability between groups are small compared with the range of individual difference between groups. Third, the purpose of our analysis was to contribute to the debate on school tracking by pointing to its possible societal implications (i.e. social segregation) and ethical issues (unequal access to knowledge).

References

Alegre, M., Ferrer, G., 2009. School regimes and education equity: some insights based on PISA 2006. British Educational Research Journal, 1–29.

Ammermüller, A., 2005. Educational opportunities and the role of institutions. ZEW Discussion Paper 44, 40.

Arneson, R. J., 1989. Equality and equal-opportunity for welfare. Philosophical Studies 56 (1), 77–93.

Barro, R., 2001. Human capital and growth. American Economic Review 91 (2), 12–17.

Barry, B., 1991. Chance, choice and justice. In: Barry, B. (Ed.), Liberty and justice: essays in political theory. Vol. 2. Oxford: Oxford University Press.

Benabou, R., 1996. Heterogeneity, stratification, and growth: Macroeconomic implications of community structure and school finance. American Economic Review 86 (3), 584–609.

Betts, J., 2011. The economics of tracking in education. In: Hanushek, E., Machin, S., Wößmann, L. (Eds.), Handbook of the Economics of Education. Amsterdam: Elsevier.

Betts, J., Roemer, J., 2005. Equalizing opportunity for racial and socioeconomic groups in the United States through educational finance reform. Department of Economics UCSD 14.

Bishop, J., 1992. The impact of academic competencies on wages, unemployment, and job performance. Carnegie-Rochester Conference Series on public policy 37, 127–194.

Bourguignon, F., Ferreira, F. H. G., Menendez, M., 2007. Inequality of opportunity in Brazil. Review of Income and Wealth 53 (4), 585–618.

Brunello, G., Checchi, D., 2007. Does school tracking affect equality of opportunity? New international evidence. Economic Policy (52), 781–861.

Checchi, D., Flabbi, L., 2007. Intergenerational mobility and schooling decisions in Germany and Italy: The impact of secondary school tracks. IZA Discussion Papers (2876).

Checchi, D., Peragine, V., Serlenga, L., 2010. Fair and unfair income inequalities in Europe. IZA Discussion Paper No. 5025.

Cohen, G. A., 1989. On the currency of egalitarian justice. Ethics 99 (4), 906–944.

Coleman, J., 1966. Equality of educational opportunity. Washington, DC: US. GPO.

Contini, D., Scagni, A., Riehl, A., 2008. Primary and secondary effects in educational attainment in Italy. LABORatorio R. Revelli Working Papers Series (82).

Davidson, R., 2009. Reliable inference for the Gini index. Journal of Econometrics 150 (1), 30–40.

de Bartolome, C. A. M., 1990. Equilibrium and inefficiency in a community model with peer group effects. Journal of Political Economy 98 (1), 110–133.

Duflo, E., Dupas, P., Kremer, M., 2009. Peer effects and the impact of tracking: Evidence from a randomized evaluation in Kenya, mimeo, MIT.

Duncan, D., Duncan, B., 1955. A methodological analysis of segregation indexes. American Sociological Review 20 (2), 210–217.

Dworkin, R., 1981a. What is equality .1. Equality of welfare. Philosophy & Public Affairs 10 (3), 185–246.

Dworkin, R., 1981b. What is equality .2. Equality of resources. Philosophy & Public Affairs 10 (4), 283–345.

Epple, D., Newlon, E., Romano, R., 2002. Ability tracking, school competition, and the distribution of educational benefits. Journal of Public Economics 83 (1), 1–48.

Epple, D., Romano, R. E., 1998. Competition between private and public schools, vouchers, and peer-group effects. American Economic Review 88 (1), 33–62.

Figlio, D. N., Page, M. E., 2002. School choice and the distributional effects of ability tracking: Does separation increase inequality? Journal of Urban Economics 51 (3), 497–514.

Fleurbaey, M., 1995. Equality and responsibility. European Economic Review 39, 683–689.

Fleurbaey, M., 2006. Capabilities, functionings and refined functionings. Journal of Human Development 7 (3), 299–310.

Fleurbaey, M., 2008a. Equal opportunity or equal social outcome? Economics and Philosophy. Cambridge Univ Press.

Fleurbaey, M., 2008b. Fairness, Responsibility, and Welfare. OUP Catalogue, Oxford University Press.

Fleurbaey, M., Peragine, V., 2009. Ex ante versus ex post equality of opportunity. ECINEQ working paper n. 141/2009.

Guyon, N., Maurin, E., McNally, S., 2010. The effect of tracking studens by ability into different schools: A natural experiment. CEPR Discussion Paper (DP7977).

Hanushek, E., Wößmann, L., 2006. Does educational tracking affect performance and inequality? Differences- in-differences evidence across countries. Economic Journal 116 (510), C63–C76.

Hanushek, E. A., Kain, J. F., Markman, J. M., Rivkin, S. G., 2003. Does peer ability affect student achievement? Journal of Applied Econometrics 18 (5), 527–544.

Hanushek, E. A., Wößmann, L., 2008. The role of cognitive skills in economic development. Journal of Economic Literature 46 (3), 607–668.

Hayfield, T., Racine, J., S., 2008. np: Nonparametric kernel smoothing methods for mixed datatypes. R package version 0.14-3.

Holmlund, H., Lindahl, M., Plug, E., 2008. The causal effect of parent's schooling on children's schooling: A comparison of estimation methods. IZA Discussion Paper Series 3630.

Hutchens, R., 2004. One measure of segregation. International Economic Review 45 (2), 555–578.

Jenkins, S. P., Micklewright, J., Schnepf, S. V., 2008. Social segregation in secondary schools: How does England compare with other countries? Oxford Review of Education 34 (1), 21–37.

Koenker, R., 2005. Quantile regression. New York: Cambridge University Press.

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

Lavy, V., Silva, O., Weinhardt, F., 2009. The good, the bad and the average: Evidence on the scale and nature of ability peer effects in school, NBER Working Paper 15600.

Lefranc, A., Pistolesi, N., Trannoy, A., 2008. Inequality of opportunities vs. inequality of outcomes: Are western societies all alike? Review of Income and Wealth 54 (4), 513–546.

Lefranc, A., Pistolesi, N., Trannoy, A., 2009. Equality of opportunity and luck: Definitions and testable conditions, with an application to income in France. Journal of Public Economics 93 (11-12), 1189–1207.

Levy, H., 1998. Stochastic Dominance: Investment Decision Making under Uncertainty. Boston: Kluwer Academic Publishers.

Li, Q., Racine, J., 2007. Nonparametric Econometrics: Theory and practice. Princeton University Press.

Li, Q., Racine, J. S., 2008. Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. Journal of Business & Economic Statistics 26 (4), 423–434.

Manski, C. F., 1993. Identification of endogenous social effects - The reflection problem. Review of Economic Studies 60 (3), 531–542.

Meghir, C., Palme, M., 2005. Educational reform, ability, and family background. American Economic Review 95 (1), 414–424.

Mullis, I., Martin, M., Kennedy, A., Foy, P., 2007. PIRLS 2006 International Report: IEA's Progress in International Reading Literacy Study in Primary Schools in 40 Countries. IEA, TIMMS and PIRLS international study center, Lynch school of education, Boston College.

Nechyba, T. J., 2000. Mobility, targeting, and private-school vouchers. American Economic Review 90 (1), 130–146.

Nickell, S., 2004. Poverty and worklessness in Britain. Economic Journal 114 (494), C1–C25.

OECD, 2005. PISA 2003 Data Analysis Manual: SAS users. OECD Programme for International Student Assessment.

OECD, 2006. PISA 2006: Science Competencies for Tomorrow's World. OECD Programme for International Student Assessment.

OECD, 2009. PISA 2006 Technical report. OECD Programme for International Student Assessment.

Pekkarinen, T., Uusitalo, R., Kerr, S., 2009. School tracking and intergenerational income mobility: Evidence from the Finnish comprehensive school reform. Journal of Public Economics 93 (7-8), 965–973.

Peragine, V., Serlenga, L., 2007. Higher education and equality of opportunity in Italy. IZA Discussion Paper Series (3163), 33.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., Rasbash, J., 1998. Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society Series B-Statistical Methodology 60, 23–40.

Pistolesi, N., 2009. Inequality of opportunity in the land of opportunities, 1968-2001. Journal of Economic Inequality 7, 411–433.

Plug, E., Vijverberg, W., 2003. Schooling, family background, and adoption: Is it nature or is it nurture? Journal of Political Economy 111 (3), 611–641.

Ramos, X., Van de Gaer, D., 2009. Empirical evidence on inequality of opportunituy.

Raudenbush, S. W., Bryk, A. S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods. Thousand Oaks, CA: Sage.

Roemer, J. E., 1993. A pragmatic theory of responsibility for the egalitarian planner. Philosophy & Public Affairs 22 (2), 146–166.

Roemer, J. E., 1998. Equality of opportunity. Cambridge: Harvard University Press.

Schutz, G., Ursprung, H. W., Wossmann, L., 2008. Education policy and equality of opportunity. Kyklos 61 (2), 279–308.

Sen, A., 1980. Equality of what? In: McMurrin, S. (Ed.), The Tanner Lectures on Human Values. Vol. 1. Salt Lake City: University of Uthh Press.

Shorrocks, A. F., 1983. Ranking income distributions. Economica 50, 3–17.

Van de gaer, D., 1993. Equality of opportunity and investment in human capital. Ph.D. thesis, K.U. Leuven.

Van de gaer, D., Vandenbossche, J., Figueroa, J., 2011. Children's health opportunities and project evaluation: Mexico's oportunidades program. Working Papers of Faculty of Economics and Business Administration, Ghent University, D/2011/7012/54.

Waldinger, F., 2007. Does ability tracking exacerbate the role of family background for student's test scores?, London School of Economics and Political Science.

Waltenberg, F. D., Vandenberghe, V., 2007. What does it take to achieve equality of opportunity in education? An empirical investigation based on Brazilian data. Economics of Education Review 26 (6), 709–723.

Wheelock, A., 1992. Crossing the Tracks. How Untracking Can Save America's Schools. New York Press.

5

An environment-adjusted evaluation of local police effectiveness: evidence from a conditional Data Envelopment Analysis approach[1]

5.1 Introduction

With institutional budgets being tight and resources being scarce, the mainly 'laissez-faire' approach towards police authorities declined considerably in the last decades. Like in other public sectors, there has been an increasing movement towards accountability. Police departments are more and more subject to performance evaluations. Good performance requires that a police

[1]This chapter is the result of joint work with Nicky Rogge (HUB).

department provides services in an efficient (at the least costs) and effective (conform the objectives) manner.

The efficiency aspect, that is, providing police services at the least costs, has been subject of multiple studies in the Operations Research literature. Several studies have embraced the DEA-framework and used it to study the efficiency of police departments, both at the level of local police districts and at higher levels (e.g., regions and countries). Examples include Drake and Simper (2002, 2003, 2004, 2005), Nyhan and Martin (1999), Sun (2002), Thanassoulis (1995), Cherchye et al. (2006) and Wu et al. (2010).

However, up to now, the effectiveness aspect of policing, i.e., providing services that fit the purposes, has remained largely unexplored by the Operations Research literature. Effective policing requires first the support and recognition of the general public (see, among others, the "police by consent" idea of Carter (2002)). Second, given the rapidly changing societies, effective policing is found to require that the police organization continuously seeks to redefine its role in the community and its relationship with the community's residents ('community policing' idea as in Beck et al. (1999)). In doing so, police officials and policy makers should ask themselves what lives in the community and what actions local police corpses could take to make police services more responsive to the needs and the expectations of citizens (that is, transform the local police organization so that more attention and resources are being dedicated to the relevant functions or activities) (Hesketh, 1992). Citizens can thus be of crucial use to identify problems in the community and provide useful feedback regarding strengths and weaknesses of community oriented police corpses. In that perspective, data on citizen satisfaction with the local police effectiveness are needed (the belief is that citizens are able to form a good impression of how good the local police corps is doing on the various functions). Note that, although citizen satisfaction measures are most commonly used, there are also other ways of measuring police effectiveness (e.g., clear up rates, internal evaluations, etc.). Nevertheless, given the particular structure of the Belgian Police (see later) with the Federal police focusing on supranational police tasks and the local police forces focusing on community-oriented policing, we believe that hard data (e.g., clear up rates) are more accurate to estimate the effectiveness of federal police organization(s), whereas

citizen satisfaction measures are more appropriate in the evaluation of local police effectiveness.

A large policing literature is devoted to measuring police effectiveness using citizen satisfaction data.[2] However, policing literature lacks a well-established evaluation methodology to compare the effectiveness of police forces that have multiple tasks and are operating in a heterogeneous environment. As noted by for instance Schafer et al. (2003, p.442), a consistent approach to measure police effectiveness based on citizen surveys has yet to emerge. Existing evaluation practices are frequently being criticized as overly simplistic and incapable of overcoming some crucial (and sensitive) issues.

One such an issue is that previous literature (e.g., Webb and Marshall, 1995; Worrall, 1999) traditionally viewed police effectiveness (and the public's satisfaction with police effectiveness) as a one-dimensional construct (that is, are citizens generally satisfied with the police services?), whereas given the multiple tasks of police, it is imperative to consider police effectiveness as multidimensional. However, developing a multidimensional measure of police effectiveness is not straightforward. One important question is how one should weight and aggregate the citizens' satisfaction with various police functions into one overall effectiveness score. This raises the question of the importance of each police task. Is it legitimate to assign equal weights to the various aspects of policing, thereby implicitly assuming that all police tasks have an equal importance? Also, is it legitimate to apply a uniform set of weights to all police departments? The knowledge that each police department has to cope with its own particular problems and specificities, seems to suggest the opposite. That is, a more flexible weighting approach, one that allows some specialization in the evaluation of the police effectiveness, is warranted.

Second, numerous studies confirm that the police do not operate in vacuum but in an open environment influenced by multiple actors and factors. As this environment is outside the control of the police, one should correct for its influences in the evaluation of police effectiveness. If not,

[2] An anonymous referee pointed out that there is also interesting literature about measuring the performance of police departments using management models. Nicholson-Crotty and O'Toole (2004), for instance, propose to apply a more formal model of public management in the evaluation of police department performances.

evaluations are very likely to be considered as unfair. Not without reason, disillusioned police departments might argue that they should be evaluated only on those aspects for which they can be held accountable and not faulted for being less effective due to less favourable operating environments. However, although several studies illustrated the impact of the operating environment on citizen satisfaction with police effectiveness (see Figure 5.2 below for a brief overview of the literature and the literature findings), the idea of actually correcting evaluations of police effectiveness based on citizen questionnaire data for environmental variables has remained largely unpursued in the literature.

This paper contributes to the literature in that it argues for a well-established Operation Research framework for evaluating police effectiveness based on citizen satisfaction data that addresses the above-stated issues. In particular, we suggest an adjusted version of the Data Envelopment Analysis (DEA) methodology for constructing scores of local police effectiveness which are multidimensional and environment-adjusted. This so-called 'Benefit-of-the-Doubt' (BoD) model (after Melyn and Moesen, 1991) exploits the key characteristic of DEA, namely that it, thanks to its linear programming formulation, allows for an endogenous weighting of the citizen satisfaction rates on multiple aspects of policing into an overall effectiveness score. We design a BoD-model (using insights from the robust and conditional order-m DEA-framework proposed by Cazals et al. (2002), Daraio and Simar (2005), (2006), (2007), (2007b), Badin et al. (2010a) and (2010b)) such that it provides multidimensional scores of police effectiveness which are (1) robust to the influences of local police departments with atypical effectiveness performances in the data (if present), (2) corrected for differences in the operating environments among police departments, and (3) allowing for non-parametric statistical inference and a visualization of the relationships between the environmental characteristics and the estimate of police effectiveness.

To illustrate the practical usefulness of the approach, we apply the model to citizen satisfaction data on Belgian local police departments. Since the wake of the thorough police reform in 1998 (the so-called Octopus Agreement signed by eight political parties and, consequentially, the Law on an Integrated Police Corps 07/12/1998), community oriented policing is top priority of local police zones. Consequently, large (financial) effort is made to construct detailed and

representative data on citizen satisfaction with the (local) police authorities. The combination of the high policy relevance and data of exceptional quality makes Belgium an interesting place to investigate effectiveness of community oriented local police corpses.

Belgian police is structured on two levels: the federal level and the local level. The federal police carry out tasks on the whole Belgian territory (they operate under the supervision of the Minister of Home Affairs and the Minister of Justice), that is, supra-local police tasks which, because of their extent, organization or consequences, cross the borders of a zone, a district or a country. Examples of such tasks are combating organized crime, drug trafficking, investigating murder cases, etc. The local police corpses operate in a local police zone, that is, a group of (small) municipalities, one (medium to large) municipality/city, or a subregion of a large city. Their task mainly consists in providing citizens with community oriented policing. It is stipulated by the Royal Decree 17/09/2001 ('*Koninklijk Besluit*' in Dutch or KB) that local police departments should carry out services and tasks that are related to the following 6 basic police functions: 'Community policing', 'Reception of citizens', 'Intervention', 'Aid to victims', 'Local investigations and detections', and 'Maintenance of public order'.[3] All six basic functions are believed to be very important to the community oriented policing and, as such, they should be considered in the effectiveness evaluations of local police departments. In this study, the focus is exclusively on the effectiveness of the community oriented local police departments.[4]

The remainder of this paper is organized as follows. In the next section, we discuss the citizen satisfaction data to measure the police effectiveness for a sample of local police departments in Belgium. Section 5.3 presents the basic BoD-methodology as well as its robust and conditional extension. In section 5.4, we present the robust estimates of the multidimensional and environment-adjusted effectiveness scores for our sample of local police departments in Belgium. Particular attention is given to how these effectiveness scores are related to a series of

[3]Only recently, the Royal Decree of 16/10/2009 added a 7th basic police function 'Traffic' (which was formerly largely included in the sixth basic function 'Maintenance of public order'). As the studied dataset only includes data from before 2009, we still employ the structure of the 6 basic police functions.

[4]For a more comprehensive presentation of the police landscape in Belgium, see also www.polfed-fedpol.be.

environmental variables characterizing the operating environment of local police departments. No doubt, this will provide information that is useful for police management and policy makers. In a final section, we make some concluding remarks and provide some directions for further research.

5.2 Data

We use data on citizen satisfaction with the local police corpses in Belgium that is collected from the Security Monitor ("Veiligheidsmonitor" in Dutch). It concerns a large-scale population survey in which several safety-related topics such as victimization, neighbourhood problems, feelings of insecurity, assessments of police contact inside and outside the context of victimization, and assessments of police effectiveness both at the federal level and the level of the municipality and/or the local police zone, etc. are treated. This Security Monitor is organized biannually by the Directorate of the National Database ('de Directie van de Nationale Gegevensbank' in Dutch) in assignment of the Minister of Home Affairs. More precisely, we use data on the citizen satisfaction with local police effectiveness as collected from the last four evaluation rounds, i.e., data from the Security Monitor administered in the years 2002, 2004, 2006, and 2008 are pooled into one dataset. In these evaluation rounds, telephone interviews were conducted in respectively 43, 64, 66, and 36 local police departments. In total 84 different police departments out of a total of 196 police departments are present in the data set, so some police departments took part multiple times in several Security Monitors. The target population were citizens of 15 years and older that resided in the examined local police zones. Random samples were drawn at the level of the individual municipality or police zone using the computer-assisted telephone interview system (CATI) with random digit dialing (from a database of fixed telephone lines) and re-weighting for respondent type (in particular, the age and the gender of the respondent).

To measure the satisfaction of citizens with local police effectiveness on the six basic functions, 21 questionnaire items are selected from the Security monitor (i.e., community policing (5 items), reception of citizens (3 items), intervention (5 items), aid to victims (1 item), local investigations

and detections (4 items), and maintenance of public order (3 items)). For an overview, we refer to Figure 5.1. All items use Likert scales to measure citizen satisfaction. Individual citizen rates on the items are aggregated at the level of the local police department (the unit of analysis) by computing the relative number of citizens that rated the performance of the local police department positively.[5] The data on the 21 items are first aggregated as relative scores at the level of the six basic police functions using a standard BoD-model (see next section).[6] The summary statistics of the relative citizen satisfaction scores at the level of the six basic police functions can be found in the upper part of Table 5.1.

[5]For a more comprehensive presentation of the 21 questionnaire items, the Likert scales used to measure citizen rates on the items and the methodology for aggregating the individual respondent perceptions at the level of the local police corpses, we refer to Rogge and Verschelde (2012).

[6]More in particular, six basic BoD-models are used to aggregate the citizen satisfaction rates on the underlying police tasks into a composite citizen satisfaction score for each basic police function (one BoD-model per basic police function). In these BoD-computations, BoD-weights are allowed to vary around the equal weights, i.e., equal weight +/- 25%. Note that the use of an arithmetic average or first principal component was also considered. However, the use of an arithmetic average requires the imposition of uniform weighting, which we try to avoid. To use the first principal component as aggregate, correlation between sub-items should be high to avoid internal inconsistency. For some basic police functions (i.e., community policing, maintenance of public order), this was clearly not the case.

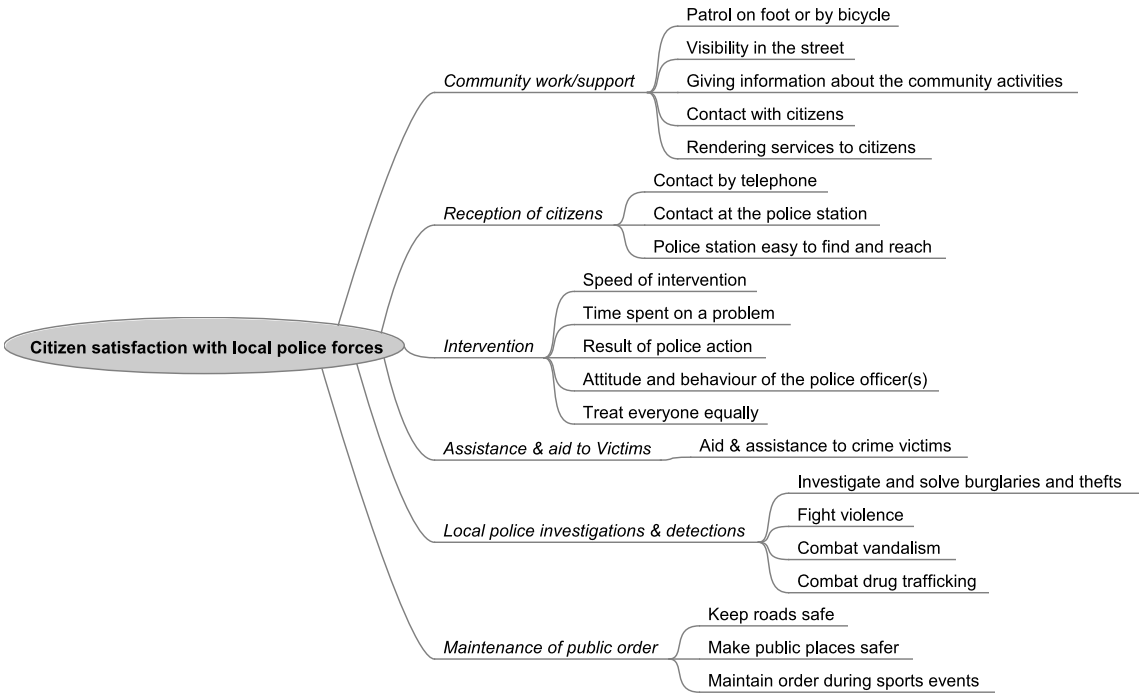


Figure 5.1: The 6 basic police functions of local police departments in Belgium

Next to the questionnaire data as collected from the Security Monitor, we also use data on environmental characteristics of the local police departments from the Directorate-general Statistics and Economic Information.[7] A literature review on the relationship between citizen satisfaction with and citizen perceptions of police effectiveness and environmental characteristics, summarized in Figure 5.2, shows the importance of demographic, socioeconomic and neighbourhoud-municipality characteristics.[8] To control for the influence of the environment, we select data on the region in which the police department is located, the year in which the data were collected, the welfare index of the local police zone, the Subsistence Income Rate, the green pressure in the police zone, and the typology of the municipality (or group of municipalities) in which the local police department is active. All of these are environmental characteristics in the sense that they

[7]Data obtained via statbel.fgov.be and aps.vlaanderen.be/lokaal/lokale_statistieken.

[8]Note that the list of environmental characteristics as in Figure 5.2 is not exhaustive. For instance, several studies also argued that the personal contact that citizens had with the police is a significant determinant of their satisfaction with the police and the police services (e.g., Schafer et al., 2003; Scaglion and Condon, 1980; Murty et al., 1990; and Webb and Marshall, 1995).

are non-controllable to the local police department but nevertheless may influence the opportunities of the local police department to operate effectively. Though not all environmental variables as in Figure 5.2 are accounted for in the subsequent analyses, the belief is that the selection of environmental characteristics captures the operation environment of local police departments in Belgium quite well. For instance, in Belgium, education and ethnicity are highly related with the observed socioeconomic variables (welfare index or Subsistence Income Rate) and demographic variables (green or grey pressure).[9]

[9]Regarding a study of the relation between the environmental characteristic 'crime rate' and police effectiveness, there are potential endogeneity problems arising from reverse causality (with the 'crime rate' in the local police zone being partially an outcome of local police effectiveness and police effectiveness being potentially influenced by citizens' perceptions of crime). Inclusion of crime rate as an environmental variable in the BoD-model is thus not advisable.

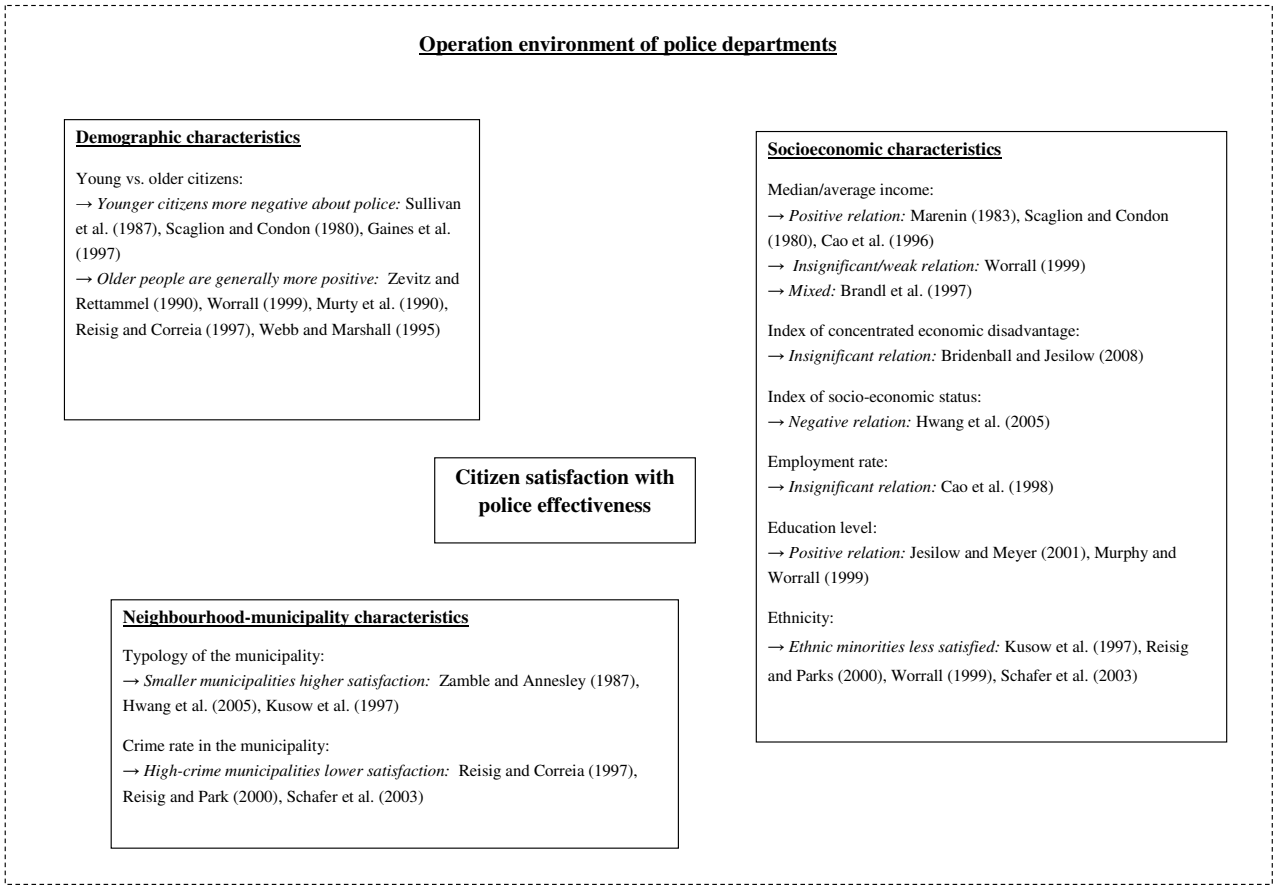


Figure 5.2: The link between the operation environment and citizen satisfaction with (local) police departments: findings of past literature

The first background characteristic 'Region' indicates whether the local police department is operating in Flanders, the Brussels region, or Wallonia (the three main regions of Belgium). The variable 'Year (02-04-06-08)' indicates whether the data are from the Security Monitor administered in 2002, 2004, 2006, or 2008. The 'Welfare index' is an important indicator of the socioeconomic status of the local police zone. It compares the average fiscal income of the citizens in a certain municipality compared to the average income of citizens in Belgium (the latter is set equal to 100). Thus, a municipality with a welfare index below (higher than) 100, resides citizens

with an average income that is lower (higher) than the income of the average citizen in Belgium. The variable 'Subsistence Income Rate' computes the percentage of citizens in the police zone with an income below the minimum standard that receive an income allowance. The variables 'Green pressure' and 'Grey pressure' are demographic indicators that measure per municipality the ratio of respectively young citizens (age 0-19 years) and older citizens (aged 60-plus) to the so-called productive population (i.e., citizens with an age between 20-59 years).

Based on the typology and the population in the municipality (municipalities) in the police zone, police zones were classified in five categories ('Typology of municipality'). According to the standard typology scheme, there are five types of municipalities: municipalities of type 1 '(large) city', municipalities of type 2 'large, regional cities and municipalities in Brussels (i.e., morphologically strongly urbanized and highly equipped municipalities)', municipalities of type 3 'metropolitan municipalities and highly equipped small cities', municipalities of type 4 'moderately to weakly equipped small city and strongly morphologically urbanized municipalities', and municipalities of type 5 'morphologically moderately and weakly urbanized municipalities'.[10,11] The summary statistics of the environmental characteristics are displayed in the lower part of Table 5.1. A visualization of the environmental characteristics 'Green pressure', 'Subsistence Income Rate', 'Welfare index', and 'Typology of municipality' for the local police departments evaluated in the Security Monitor 2006 can be found in Figure 5.6 in Appendix.

[10]The typology of the municipalities was designed by the General Police Support Service and the Police Service Policy Support (Algemene Politie Steundienst and dienst Politiebeleidsondersteuning, or APSD/PBO, in Dutch) and is used in various police statistics since 1996.

[11]For police zones with more than one municipality, the municipality with the highest level of urbanization determines the category for the police zone (provided that more than 35% of the inhabitants of the police zone are resident in that municipality).

Variable	Mean	St.Dev.	Min	Q1	Med	Q3	Max
Output							
Community policing	0.837	0.096	0.608	0.755	0.850	0.915	1.000
Reception of citizens	0.862	0.076	0.664	0.796	0.875	0.926	1.000
Intervention	0.806	0.080	0.607	0.744	0.814	0.861	1.000
Aid to victims	0.866	0.059	0.688	0.826	0.876	0.906	1.000
Local investigation	0.882	0.063	0.701	0.833	0.889	0.927	1.000
Maintenance of public order	0.925	0.041	0.788	0.898	0.929	0.954	1.000
Environmental variables							
Subsistence Income Rate ($\times$ 100)	0.869	0.727	0.099	0.310	0.659	1.160	3.574
Green Pressure	42.077	4.413	32.474	38.455	42.572	44.817	53.763
Grey Pressure	41.624	7.825	24.762	38.054	40.615	44.376	75.145
Welfare Index	98.317	12.642	70.638	87.716	100.159	107.534	138.00
Variable	Groups	Mean size	St.Dev	Min	Med	Max	
Typology	5	41.80	24.87	19	32	69	
Region	3	69.67	49.01	20	71	118	
Year	4	52.25	15.02	36	53.50	66	
Observations	209						

Table 5.1: Summary statistics for the local police departments

5.3 Methodology

5.3.1 The 'Benefit-of-the-Doubt' (BoD) model

To estimate the multidimensional measure of local police effectiveness based on the citizen satisfaction rates, we advocate a construction methodology that is rooted in the popular DEA-method. This DEA-method is a non-parametric efficiency measurement technique originally developed by Farrell (1957) and put into practice by Charnes et al. (1978), to measure the relative efficiency performance of a set of similar entities (organizations, production lines, local police departments, etc.) which employ (possibly) multiple inputs to produce (possibly) multiple outputs in complex operating settings typically characterized by no information on the prices of inputs and outputs

and/or no (exact) knowledge about the 'functional form' of the production or cost function.[12]

The custom made version of the DEA-model, the so-called BoD-model (after Melyn and Moesen, 1991), that is used here to construct a multidimensional measure of local police effectiveness differs from the traditional DEA-model in that it only looks at the output dimension without explicitly taking into account the input dimension. Formally, in the DEA-setting, all evaluated entities (i.e., local police departments) are assumed to have a 'dummy input' equal to one.[13] The application at hand consists of six outputs, i.e., the six basic functions of local police departments as discussed in the previous section.

The conceptual starting point of the BoD-model is that in the aggregation of the outputs into one composite output score, in the absence of detailed information on the true weights for the outputs, information on the weights can be retrieved from the observed data themselves. The BoD-model determines the weights for the outputs endogenously by looking a priori at the observed performance data. More precisely, the basic idea of the BoD-model is to put the data of the evaluated entity in relative perspective to the performance data of all entities in the sample set, and look for the outputs of relative strength and of relative weakness.

The notion of the 'Benefit-of-the-Doubt' enters into the interpretations of the relative performances and the specification of the weights that follow from these interpretations. Particularly, basic police functions on which the evaluated local police department performs relatively well (i.e., a relatively high number of citizens rating the effectiveness of the evaluated local police department positively) are interpreted as basic police functions in which the local police department is relatively good (i.e., a relative strength in the functioning of that department) or as basic police functions which are considered to be relatively more important by that department (thus, with the department assigning more time, resources, and effort to it). Given this, the effectiveness realized in these basic police functions should weigh more heavily in the evaluated department's overall effectiveness score. Therefore, the BoD-model assigns a high endogenous weight to such basic

[12] For an extensive overview of the DEA literature, we refer to Simar and Wilson (2008).

[13] See Lovell and Pastor (1999) for an extensive discussion on DEA models without inputs or without outputs.

police functions. The opposite reasoning holds in the interpretation of basic police functions on which only a relatively small number of citizens rated the effectiveness of the evaluated local police department positively. In essence, this means that the BoD-model grants each local police department the benefit-of-the-doubt when it comes to assigning weights in the composition of its score of overall effectiveness. The resulting BoD-weights $w_{c,i}$ are chosen in such a way that the evaluated department's effectiveness score E_c is maximized. In formal notations:

$$E_c = \max_{w_{c,1},\ldots,w_{c,q}} \sum_{i=1}^{q} w_{c,i} y_{c,i} \tag{1}$$

$$s.t.$$

$$\sum_{i=1}^{q} w_{c,i} y_{j,i} \leq 1 \qquad \forall j = 1,\ldots,c,\ldots,n \tag{1a}$$

$$w_{c,i} \geq 0 \qquad \forall i = 1,\ldots,q, \tag{1b}$$

with n the number of local police departments in the dataset Υ (i.e., n=209); E_c the BoD-estimated score of local police effectiveness for the local police department c; q the number of basic police functions on which the local police departments are evaluated (here, q=6); $y_{c,i}$ the citizen satisfaction score of police department c on the basic police function i; $y_{j,i}$ the citizen satisfaction score of police department j $(j = 1,\ldots,c,\ldots,n)$ on the basic police task i; and $w_{c,i}$ the optimal BoD-weight assigned to the basic police function i for the local police department c under evaluation.

Note the two constraints in the BoD-model. Restriction (1a) is a normalization constraint which imposes that when applying the optimal BoD-weights of the evaluated local police department to all other departments in the sample set Υ, the overall effectiveness scores of all departments should be smaller than or equal to one. Thus, it holds that $0 \leq E_c \leq 1$. In the interpretation of the effectiveness scores E_c, higher scores indicate better relative effectiveness performances. In addition, when the evaluated local police department is evaluated with $E_c < 1$, this indicates that there is at least one other police department in the sample set Υ that realizes a better overall effectiveness score even when applying the evaluated police department's most favourable

weights $w_{c,i}$. In other words, based on the observed performances in the dataset, there is still room for improvement. If the evaluated local police department obtains the maximal score of one (i.e.,$E_c = 1$), it is not outperformed by other departments in the dataset when applying the own best possible weights $w_{c,i}$. That is, the evaluated police department is indicated as its own benchmark. The non-negativity constraint (1b) limits the optimal weights $w_{c,i}$ to be non-negative. Consequently, an increase in the citizens rating of the local police department on a particular basic police function, *ceteris paribus*, will not result in a lower effectiveness score.

Admittedly, some may criticize the large flexibility in basic BoD-weighting since it could possibly lead to unfortunate and/or misleading evaluation findings. This criticism is not completely unfounded: the basic, unrestricted BoD-model as in (1)-(1b) may assign zero weights and/or unrealistically high weights to one or multiple basic functionalities without violating the two aforementioned restrictions. As such, the basic BoD-model can ignore and/or overemphasize one or more of the basic police functions in the composition of the overall effectiveness score E_c (thus allowing for a too high (undesirable) degree of "specialization" in the effectiveness evaluations of the local police departments).

This problem of improper optimal BoD-weights has already been discussed extensively in the literature (see Thanassoulis et al. (2004) and Cherchye et al. (2007) for an elaborate discussion of this topic). It has been argued that the problem can be largely alleviated by consulting a group of stakeholders (e.g., the interviewed police officers, police chiefs, etc.) on what they believe are proper values for the weights, and incorporating their opinions into the BoD-model by adding weight restrictions. The idea is then to enforce the installation of proper weights and let subsidiarity in BoD-weighting only play within the confines set by the stakeholders.

With an eye towards practical usage, we consulted the parties most involved in the process of local policing, the chiefs of the local police departments, to gain knowledge on what they think are appropriate importance weights for the six basic functionalities of local policing. The police chiefs of a large majority of the local police departments in Belgium were consulted by email

and requested to fill out a questionnaire.[14] Each police chief was asked to distribute a total of 100 points over the six basic functionalities of local policing, thereby allocating more points to the functionalities which he/she regards as most important. A total of 63 police chiefs participated in the study (approximately 1/3 of the contacted police chiefs). Summary information about the weights so obtained (i.e., average, minimum, and maximum weights) is provided in Table 5.2.[15]

To integrate information on the opinions of the police chiefs of the local police departments into the BoD-model, we opted for using proportional virtual weight restrictions.[16] This type of weight constraint imposes that the BoD-model can choose the optimal importance of the policing functions freely within a range specified by a lower bound value α_i and an upper bound value β_i. Formally, this involves adding the following weight constraints to the standard BoD-model:

$$\alpha_i \leq \frac{w_{c,i} y_{c,i}}{\sum_{i=1}^{q} w_{c,i} y_{c,i}} \leq \beta_i \quad \forall i = 1, ..., q. \tag{1c}$$

In the application below, we set the lower bound and upper bound value equal to the 5%-percentile and 95%-percentile of weights specified by the consulted police chiefs. For instance, for the basic police function 'Community work' this involves setting $\alpha_i = 0.10$ and $\beta_i = 0.30$.[17]

[14] Another possibility would be to consult the citizens and include their opinions on the appropriate importance for the police tasks (see, for instance, Webb and Katz, 1997).

[15] For a more comprehensive discussion of the procedure used to collect the opinions of the police chiefs (as well as a presentation of the collected individual opinions), we refer to Rogge and Verschelde (2012).

[16] Note however that other types of weight restrictions have been proposed in the DEA/BoD-literature (for an overview, see, among others, Thanassoulis et al. (2004) and Cherchye et al. (2007)).

[17] Results can be sensitive to the choice of the upper and lower bounds as these bounds determine the flexibility in weight choice. As some outlying opinions determine the minimum and maximum level of weights, we used the more robust 5%-percentile and 95%-percentile of weights as constraints. Results do not alter considerably when we take the 10%-percentile and 90%-percentile of weights as bounds.

	Comm. pol.	Reception	Intervention	Aid	Ivestig.	Public order
Average	0.189	0.146	0.245	0.129	0.160	0.131
St.Dev.	0.065	0.042	0.098	0.051	0.051	0.095
Min.	0.030	0.030	0.100	0.020	0.010	0.020
5% perc.	0.100	0.082	0.160	0.030	0.050	0.050
25% perc.	0.150	0.100	0.180	0.100	0.150	0.090
Median	0.200	0.150	0.200	0.150	0.170	0.110
75% perc.	0.210	0.170	0.300	0.170	0.200	0.160
95% perc.	0.300	0.200	0.400	0.200	0.220	0.200
Max.	0.400	0.250	0.700	0.250	0.250	0.650

Table 5.2: Summary of weights specified by the consulted police chiefs

5.3.2 The robust and conditional BoD-model

The BoD-model as in (1)-(1c) still suffers from two important drawbacks. Firstly, due to the deterministic nature of the BoD-model, estimated scores of local police effectiveness are sensitive to the influences of outliers. The practical implications of this drawback can be far-reaching. Evaluated local police departments are naturally sensitive about being compared with the performances of other corpses (unless they compare well, of course). This concern is particularly acute when there is the danger of being compared against departments with outstanding evaluation outcomes due to other reasons than a high police effectiveness (as measured by citizen satisfaction in the community). Secondly, estimated effectiveness scores are not corrected for differences in the operating environments of the local police departments. As discussed in the introductory section, both the academic literature and the experiences of the local police men and women indicate that the operation environment can considerably influence the local police departments' opportunities to function in an effective manner (and, thus, to realize a relatively high E_c). Using insights of Cazals et al. (2002), Daraio and Simar (2005, 2006, 2007a,b), Badin et al. (2010a) and (2010b), we tailor the BoD-model such that it no longer suffers from these limitations. We proceed in two steps.

In a first step, we adjust the BoD-model so as to make it robust to the influences of local police departments with atypical performances in the data (if present in the sample set). To do so, we use the insights of the order-m DEA approach of Cazals et al. (2002).[18] The essential idea of this approach is to not consider the full sample set of n police departments in the definition of the effectiveness scores E_c, as in the traditional BoD-computations. Instead, under a simple bootstrapping framework, $B(b = 1, ..., B)$ computation rounds are performed (with B a large number, in casu 500), in each of which a sub sample $\Upsilon_c^{m,b}$ of only m observations (randomly and i.i.d. drawn from the full sample of n local police departments) are used in the estimation of the overall effectiveness score $E_c^{m,b}$. The robust BoD-method thus estimates B effectiveness scores $E_c^{m,b}$ by means of the linear programming problem in model (1)-(1c) after replacing Υ by $\Upsilon_c^{m,b}$. Having obtained the B effectiveness scores $E_c^{m,b}$, we compute the robust BoD-based local police effectiveness score E_c^m as the arithmetic average of these B scores. As local police departments with performance data that are atypical do not form part of the sub sample $\Upsilon_c^{m,b}$ in every draw, the impact of such departments on the order-m effectiveness scores E_c^m is effectively mitigated. In short, the order-m BoD-based effectiveness score is the benevolently computed effectiveness score of the evaluated department relative to the expected maximum effectiveness observed among m randomly drawn departments.[19]

To correct the estimate of local police effectiveness for differences in the operation environments of local police departments, in a second step, the order-m BoD-model is further extended with insights after Daraio and Simar (2005, 2006, 2007a,b). Specifically, these authors propose a methodology that obtains so-called conditional evaluation measures, which condition the perfor-

[18] See Daouia and Ruiz-Gazen (2006) and Daouia and Gijbels (2011) for theoretical and monte carlo evidence of the robust properties of the Cazals et al. (2002) partial frontier approach.

[19] In analogy with the order-m efficiency concept of Cazals et al. (2002), a less extreme benchmark is chosen to reduce the sensitivity of effectiveness estimates to the influence of outlying observations. Consequently, the order-m BoD effectiveness score (effectiveness of evaluated department relative to the expected maximum effectiveness of m randomly drawn departments) is higher than or equal to the basic (non rubust) BoD-based effectiveness score (effectiveness of evaluated department relative to the expected maximum effectiveness of all sampled departments), but converges to the basic BoD-based effectiveness score when m goes to n.

mance evaluation on exogenous factors, which we capture by the vector Z. The computation of these conditional measures involves a slight modification of the robust order-m procedure outlined above. In particular, whereas in the unconditional robust order-m procedure, in each draw, all local police departments have an equal probability of being selected for membership in the sub sample (that is, local police departments are drawn from Υ with uniform probability), in the conditional order-m framework, the probability for a local police department of being drawn is defined on the basis of a kernel density function evaluated at the location of the exogenous factors for the evaluated local police department c (see Appendix for technical details on the construction and use of appropriate kernel density weights).[20] The idea is that local police departments get a greater probability of being drawn for membership in the sub sample (label the sub samples $\Upsilon_c^{m,z,b}$) if their operation environment (as characterized by the environmental characteristics as described in Section 5.2) is more similar to the one of the evaluated local police department c. Local police departments active in operating environments that are largely dissimilar to the operating environment of the local police department under evaluation have a low, and in some cases, even no probability of being selected for membership in the sub sample (for instance, for the categorical variable 'region', we find it is optimal to give no weight to local police departments from another region). Intuitively, one could say that the conditional effectiveness measurement accounts for the operational environment by comparing likes with likes. We denote the estimates of the conditional BoD-model by $E_c^{m,z}$.

For completeness, we note that there are also other approaches to account for the operating environment in evaluations. One such approach is the frontier separation approach (e.g., Charnes et al. (1981), Portela and Thanassoulis (2001) and De Witte and Rogge (2010)). In essence, this approach consists in splitting up the complete sample of local police departments into separate comparison groups specific to a particular type of operating environment. Though being intuitively appealing, we do not use this approach as it suffers from some important limitations. One particular problem is that as the number of combinations for environmental characteristics increases, the sample size of the subgroups becomes smaller and smaller. This makes the

[20] Jeong et al. (2010) show the asymptotic properties of the conditional frontier approaches.

separation approach problematic and sometimes even infeasible. Another problem is that the splitting up approach is difficult to apply in evaluations in which the operating environment is characterized by, among other things, continuous environmental variables such as green pressure or subsistence income rate (or one should categorize these variables, e.g., green pressure $\leq 20\%, 20\% <$ green pressure $\leq 50\%$, etc., which causes loss of information and which makes that one should make arbitrary choices in the categorization).

Note that $E_c^{m,z}$ can be larger than unity. Indeed, thanks to drawing a subsample of m observations with replacement from the full sample Υ, the evaluated local police department c will not always be part of the sub sample $\Upsilon_c^{m,z,b}$. As such, "super-effective" performances (i.e., local police departments with a $E_c^{m,z}$ score higher than 1) could arise. The "super-effective" $E_c^{m,z}$ score is interpreted as a local police department that is doing better than the average m other local police departments in its reference sample (police departments that operate under largely similar environmental conditions).

We conclude this section with two remarks. First, an important parameter in both the unconditional and conditional order-m procedure is the parameter m (i.e., the number of observations against which the effectiveness of the evaluated local police department should be compared). There is no standard methodology which allows computing the most appropriate value for m. However, as pointed out by Cazals et al. (2002) and Daraio and Simar (2007b), too high and too low values of m should be avoided (for a more comprehensive discussion of the role of the parameter m, we refer to these studies). In our application, we use $m = 50$. However, sensitivity analysis points out that the results are robust with respect to alternative choices of value of m (i.e. we also considered $m = 20, 30, 40, 60, 70, 80, 90, 100$). Second, because of the re-sampling procedure, we can construct confidence intervals and standard deviations for $E_c^{m,z}$.

5.3.3 Statistical inference and visualization

As a major advantage, the conditional and robust BoD-framework allows for an interpretation of the association between the environmental characteristics Z and effectiveness of local police

departments. In particular, by non-parametrically regressing the ratio of the unconditional [i.e., without accounting for the operating environment; E_c^m] to the conditional [i.e., accounting for the heterogeneity; $E_c^{m,z}$] order-m estimates on the environmental characteristics Z, we can learn (1) whether Z is on average statistically significantly related to the overall performance scores E_c^m, and (2) whether this relationship is positive or negative. Daraio and Simar (2005, 2007a) also showed how the conditional order-m approach allows one to visualize these estimated relationships. When Z is univariate, the visualization is clear-cut (i.e., a scatter plot with on the horizontal axis the environmental and on the vertical axis the ratio $E_c^m/E_c^{m,z}$). When Z is multivariate (as in our application), the visualization is more demanding. However, partial regression plots (see Daraio and Simar (2007b), Badin et al. (2010a) and (2010b)), where only one environmental characteristic is allowed to vary while all other environmental characteristics are kept at a fixed value (ceteris paribus) provide an appealing solution. The interpretation of the results is that positive (negative) estimated regression coefficients and slopes in the visualizations indicate environmental variables that are positively (negatively) related to the overall police effectiveness. The intuition behind the Badin et al. (2010b) subsample approach is explained in Appendix.

5.4 Empirical results

Before estimating the robust and environment-adjusted BoD-based estimates of the scores of local police effectiveness, we examine the traditional version of scores of police effectiveness, that is, a one-dimensional measure of police effectiveness as measured by the global satisfaction of the citizens with the local police (i.e., based on the rates given by the respondents on one global question in the Security Monitor that measures the citizens overall satisfaction with the police and the police services). The results are presented in Table 5.3. To be comparable with the BoD-estimated effectiveness scores, we also divided the scores by its maximum to obtain relative scores in stead of absolute levels of satisfaction. On average, 86.5% of the respondents have a positive to very positive perception (high to very high satisfaction) of the police and the police services. In other words, most Belgians hold favourable impressions about their local police. Nevertheless, the difference between the minimum value of 68.5% and the maximum

value of 95.3% indicates that there is variation between local police departments. Similar results of citizens being, in general, satisfied to very satisfied with the local police were found by most other studies in the literature (e.g., Bridenball and Jesilow, 2008; Sims et al., 2002). The relative scores of the 1-dimensional citizen satisfaction score indicate that the median police zone should be able to increase its satisfaction score by 7%.

	Average	St.Dev	Min	Q1	Median	Q3	Max
1-dim. satisfaction score (absolute)	0.865	0.059	0.685	0.822	0.882	0.907	0.953
1-dim. satisfaction score (relative)	0.908	0.062	0.719	0.863	0.926	0.952	1.000
BoD-score	0.897	0.061	0.734	0.846	0.905	0.946	1.000
Conditional BoD	0.931	0.045	0.792	0.899	0.938	0.964	1.002

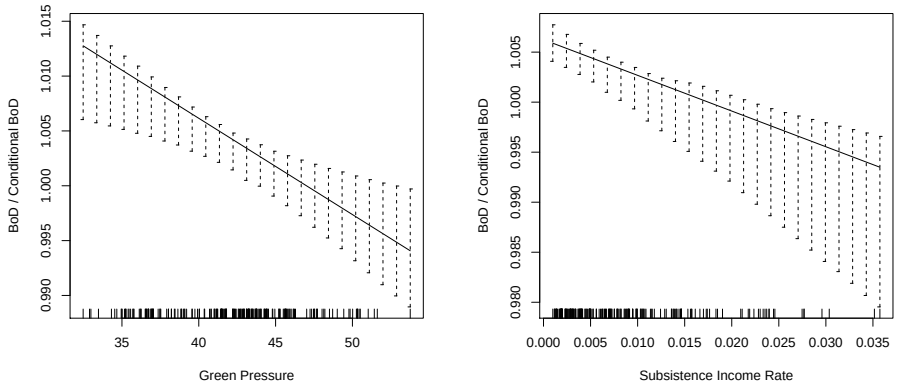
Table 5.3: Estimates of local police effectiveness in different model specifications (n=209 local police forces)

The third row of Table 5.3 presents the summary statistics of the BoD-based estimates of local police effectiveness for the local police departments, however, without any robustification or correction for differences in the operating environment (that is, the scores as computed by the BoD-model described in Section 3.1). The BoD-model evaluates four local police departments as perfectly effective (i.e., $E_c = 1$). The median score of 0.905 is rather high and indicates that the local police departments in the sample set are rather effective in terms of fulfilling their six basic police functions (as perceived by the interviewed citizens). Nevertheless, this score also indicates that there is still some room for further improvement.[21] Sensitivity tests in Appendix show that our results still hold when we relax the independence assumption in the used routines.

[21]It is correctly noted by an anonymous referee that there is also other interesting output of the BoD-model. For instance, the BoD-estimated weights for the basic police functions are especially relevant for the local police managers/chiefs as a detailed analysis of these weights (and, in particular, whether or not the weight restrictions are binding for certain basic police functions) indicates what basic police functions require additional attention. For a discussion of these and other BoD-model outcomes, we refer to Rogge and Verschelde (2012).

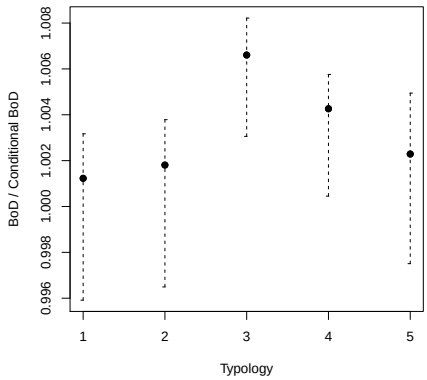
More interesting than the traditional one-dimensional and the basic BoD-based estimates are the robust and environment-adjusted measures of local police effectiveness as computed by the robust and conditional order-*m* version of the BoD-model (see section 5.3.2). We estimate a conditional BoD-model that includes the demographic variable 'green pressure', the neighbourhood characteristic 'typology of the municipality', and the socioeconomic characteristic 'subsistence income rate'. Our model also controls for the region in which the police department is operational (Flanders, Brussels, or Wallonia) and the year in which the citizen survey was administered (i.e., 2002, 2004, 2006, or 2008).

We notice that when accounting for the differences in the operating environments among the local police departments (as characterized by the selection of environmental variables), the median local police effectiveness score increases to 0.938. Thus even after the adjustment for environmental differences, for half of the local police departments, there is still room for an improvement of more than 6%. The quartile of lower-performers can increase their effectiveness by more than 10%. The lowest performer should be able to increase its effectiveness score by 20%.
Sensitivity tests, given in Appendix, show that replacing 'green pressure' for 'grey pressure' as demographic background characteristic or switching 'subsistence income rate' for 'welfare index' as socioeconomic environmental characteristic in the estimations does not alter the results considerably.

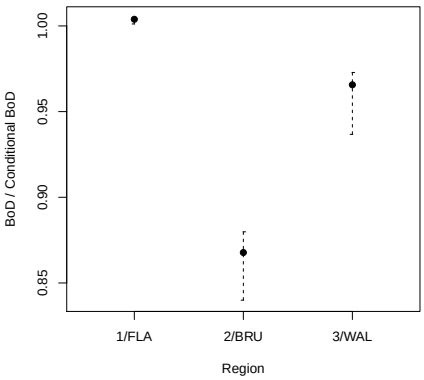


(a) Green Pressure

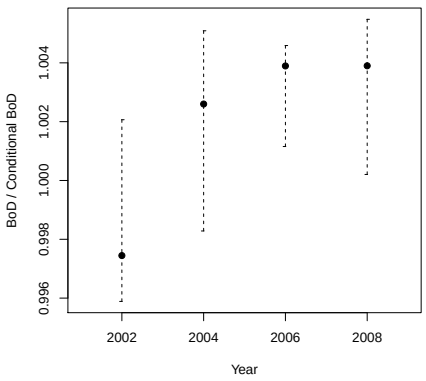
(b) Subsistence Income Rate



(c) Typology



(d) Region



(e) Year

Figure 5.3: Visualization of the results

As a first class of environmental characteristics, consider the estimated relationship between the effectiveness score for local police departments and the demographic variable 'green pressure' (see Figure 5.3(a)). The plot shows that the variable 'green pressure' is related negatively to the police effectiveness as measured by citizen satisfaction. In other words, police departments active in municipalities with a rather young population receive lower effectiveness rates. These findings are consistent with previous research (e.g., Sullivan et al., 1987; Gaines et al., 1997; Zevitz and Rettammel, 1990; and Worrall, 1999), which indicated that elder citizens are more likely to be satisfied (or very satisfied) with the local police and the local police services compared to younger citizens.

As a second class of environmental characteristics, consider the estimated relationships between the effectiveness of the local police departments and the socioeconomic variable 'subsistence income rate'. Figure 5.3(b) shows a negative relationship between police effectiveness and 'subsistence income rate'. Recall that 'subsistence income rate' can be interpreted as a measure of disadvantage in the sense that a higher value should be seen as negative. The finding of a significant effect of 'subsistence income rate' is in contrast with what was found in other studies, with, for instance, Bridenball and Jesilow (2008) showing that an index of concentrated economic disadvantage was not related to the citizen satisfaction with the police once other environmental characteristics were accounted for in the models.

As a third class of environmental characteristics, consider the estimated relationship between the typology of the municipality and the overall effectiveness score of local police departments. Figure 5.3(c) shows that local police departments in municipalities of typology 3 (i.e., metropolitan municipalities and highly equipped small cities) are rated more positively by citizens compared to their counterparts situated in municipalities of more urbanized types. This is in line with other studies that citizens living in urban police zones (i.e., typology 1 and 2) are typically less positive with the police and the police services. However, the difference between typologies are not significant at the 5% significance level.

Two other environmental characteristics for which a correction was performed in the estimations

of the effectiveness scores of local police departments are the region in which the police department is operational and the year in which the citizen perception data were collected. Figure 5.3(d) shows that local police departments that are operational in municipalities in Flanders are rated more positively by citizens in terms of police effectiveness compared to local police department that are located in the Brussels and the Wallonian region.[22] In addition, results show that it is more difficult to obtain high police effectiveness scores in Brussels than in the Wallonian region (i.e., citizens living in municipalities in the Brussels region appear to be less satisfied with the local police). Bandwidth sizes of the variable 'Region' are very close to 0 (see Table 5.6 in Appendix). This means that observations are only compared to observations from the same region. In other words, the estimation procedure points out that the 3 regions have an operating environment which is not comparable.

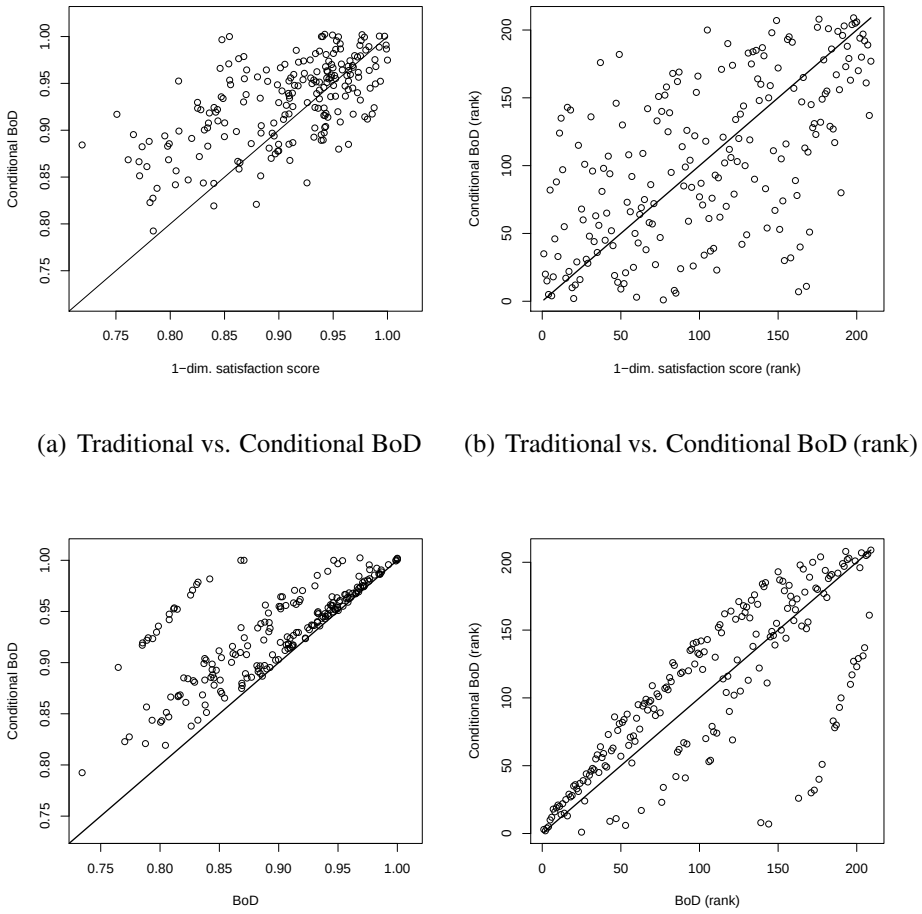
Regarding the impact of the year in which the citizen surveys were collected, Figure 5.3(e) indicates that there was an increasing trend in the effectiveness scores realized by local police departments in the period 2002-2004-2006, with particularly in the year 2006 higher average effectiveness scores for local police departments based on citizen satisfaction. The trend stopped in the year 2008 (i.e., the dots for the years 2006 and 2008 are somewhat at the same position in the plot). These finding are not really a surprise as the official police reports of Van Den Bogaerde et al. (2007) and (2009) already noted this trend. However, the large confidence intervals and, in particular, the overlap between the confidence intervals for the periods, suggests that there is no statistically significant difference between the overall effectiveness scores of the local police departments for the consecutive periods.

Figure 5.4 demonstrates how the robust and environment-adjusted BoD-generated effectiveness scores (rankings) differ from the traditional police effectiveness scores and the police effectiveness scores as estimated by the basic BoD-model (see Section 5.3.1). In particular, Figure 5.4(a) and 5.4(b) look at how the robust and environment-adjusted local police effectiveness scores (ranks) relate to the traditional effectiveness scores as measured by one global question. Given

[22]Note however that we do not claim that our sample of local police zones is representative at the regional level. The variable 'region' is included as control variable.

that correlations are rather low (i.e., Spearman rank correlation of 0.613), the evaluation outcomes seem to differ considerably. In fact, there are some local police departments that obtain low effectiveness scores (ranks) when using the traditional, unidimensional measure of police effectiveness and high effectiveness scores (ranks) when employing the conditional and robust BoD-model in the estimation of a multidimensional measure of local police effectiveness, and vice versa.

Figure 5.4(c) and 5.4(d) look at the association between the police effectiveness scores (ranks) as estimated by the BoD-model with and without a robustification and correction for differences in the operating environment of local police departments. By robustification and conditioning on the environment, effectiveness scores increase as we control for the influence of atypically good performing police departments and in general 'unfavourable' environmental variables. The plots show a high overall association (i.e., Spearman rank correlation of 0.765). However, high correlations between scores (rankings) do not imply that scores (rankings) are completely equivalent. Quite the contrary, as indicated by both the scatter plots, effectiveness scores and ranks of individual local police departments clearly depend on whether or not there was a robustification and correction for differences in the operation environments of the local police departments. The norm in effectiveness evaluations of local police departments should thus not only be on differential, 'Benefit-of-the-Doubt' weighting, but also on making the scores robust to outliers and correcting the scores for differences in the operating environments of the local police departments (as represented by the selection of environmental characteristics).



(a) Traditional vs. Conditional BoD (b) Traditional vs. Conditional BoD (rank)

(c) BoD vs. Conditional BoD (d) BoD vs. Conditional BoD (rank)

Figure 5.4: Comparison of approaches by scatter plot of scores and ranks

The practicality of the conditional BoD-approach is illustrated by focusing on the police department with largest difference between the unidimensional, unconditional satisfaction score and the conditional BoD-score. The local police department, situated in the Walloon region, is characterized by an environment of relatively many youngsters (79-percentile) and beneficiaries of a subsistence income (62-percentile). Figure 5.5 illustrates that while the department is estimated to be able to perform over 25% better by the unidimensional satisfaction score, the conditional BoD-score indicates room for improvement of 12%. In other words, by taking into account

the multidimensionality of local policing, possible outliers and the environment, the room for improvement drops by 50% for this particular department. The unconditional BoD-model indicates room for improvement of 18%, which is more than 50% higher than the conditional model. Differently put, by using the conditional BoD-model, we can control for the less favourable operating environment of the local police department.

Additionally, Figure 5.5 illustrates the estimated optimal importance of police tasks.[23] When we compare "likes with likes", we note that it is optimal to give as much weight as is possible within the restrictions to 'intervention' and 'maintenance of public order'. As discussed in detail in Rogge and Verschelde (2012), police tasks with weights that are bounded by the respective lower and upper bound weight constraints reveal respectively 'relative weaknesses' and 'relative strengths' of local police departments. In this particular case, 'intervention' and 'maintenance of public order' are indicated to be 'relative strengths' and 'community work and support' is considered as a 'relative weakness'.

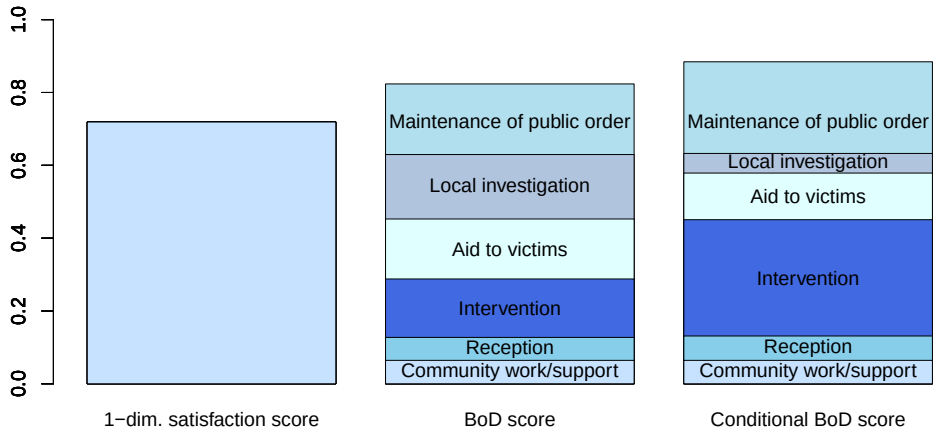


Figure 5.5: Example of the practicality of the conditional BoD approach

[23]For the conditional BoD-model, we used the bootstrap mean of the optimal importance weights.

5.5 Concluding remarks

In this paper, we advocated the use of a custom made OR framework to evaluate police effectiveness of community oriented local police forces. In particular, we suggested an adjusted version of the Data Envelopment Analysis (DEA) methodology. This 'Benefit-of-the-Doubt' (BoD) model (after Melyn and Moesen, 1991) allows the construction of a perceived effectiveness score by endogenously weighting the citizen satisfaction with the multiple aspects of policing. For each local police department, weights for the basic police functions are chosen such that the highest overall police effectiveness score is realized. Eventually, opinions of stakeholders such as police chiefs of local departments can be taken into account in the weighting. To make the BoD-based effectiveness score robust to outliers as well as corrected for differences in the operating environment, we extended the BoD-model using insights from the robust and conditional order-m DEA-framework. A major advantage of these extensions is also that they allow for non-parametric statistical inference and a visualization of the relationships between the environmental characteristics and the estimate of police effectiveness.

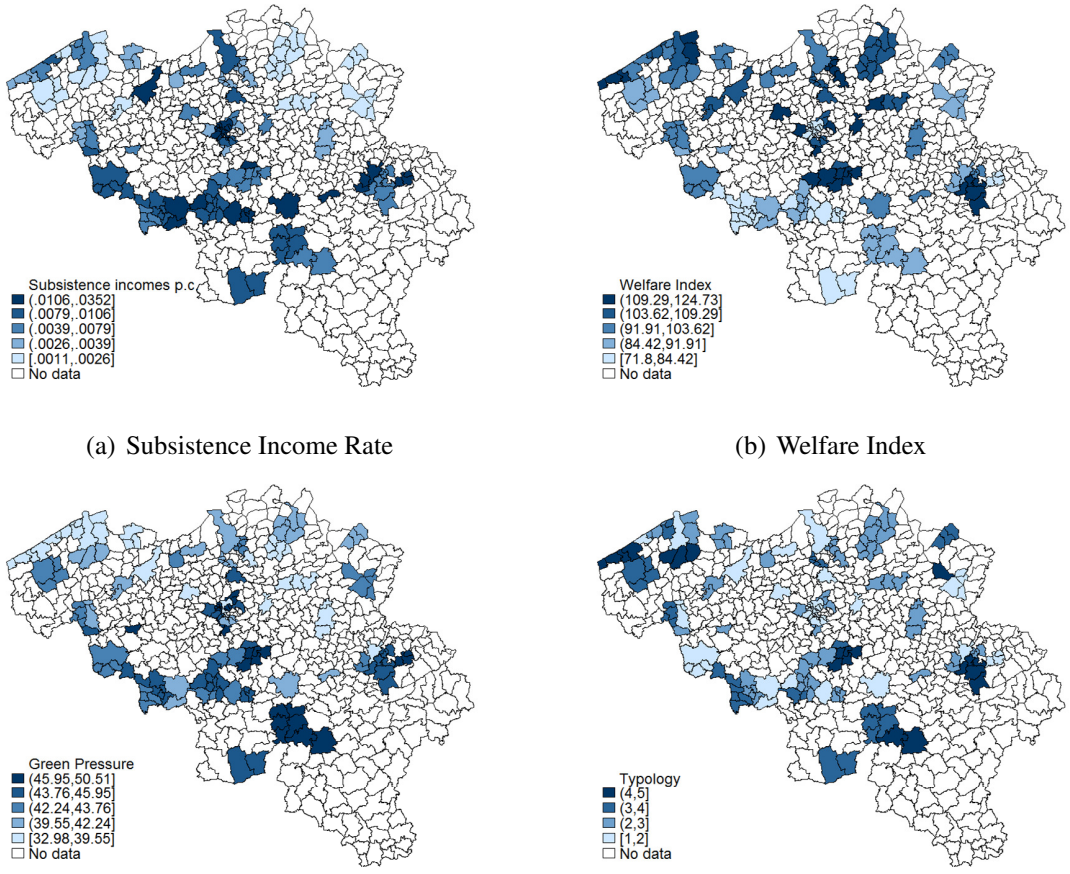
To illustrate the practical usefulness of the approach, we applied the robust and conditional BoD-model to citizen satisfaction data on Belgian local police departments. We first show that the median police effectiveness score equals approximately 0.94. This means, that, for most police departments, there is still room for further improvement.
The estimations of the relationships between the effectiveness scores and the environmental characteristics reveal that the proportion of young citizens is negatively related to the citizen satisfaction with police effectiveness. Regarding the percentage of the resided population in the local police zone that is beneficiary of a subsistence income, results point out respectively a strong negative relationship with the effectiveness of local police departments. Results are controlled for effects of urbanization, year effects and regional differences. We believe that all this information is useful to help policy makers better understand the environmental factors that influence the citizen satisfaction with and attitudes towards the police and police services.

However, there are some important reasons why the results found in this paper should be inter-

preted with caution. First of all, we emphasize that one should not generalize the results found here to other countries because the significance of the link between environmental characteristics and police effectiveness varies without doubt with the particular conditions. However, we believe it to be interesting to apply the proposed framework in other settings or to data of previous studies to check for recurrent patterns. Second, there is a risk of omitted variable bias. As a suggestion for future studies, it would be interesting to expand the selection of environmental characteristics.

5.6 Appendix

5.6.1 Visualization of the environmental characteristics of local police departments



(a) Subsistence Income Rate

(b) Welfare Index

(c) Green pressure

(d) Typology

Figure 5.6: Environmental characteristics of local police departments

5.6.2 Technical details

Conditional BoD score

The basic idea of the conditional BoD-model is that local police departments with a similar operating environment as police department c get a greater probability of being drawn for mem-

bership in the subsample of benchmark observations. We follow Badin et al. (2010a) in using the conditional distribution function $F(Y|Z=z)$ as starting point to determine the probabilities to be drawn.[24] This approach has as main advantage that no separability assumption is imposed as $F(Y|Z=z)$ captures the effect of Z on both the attainable set as on the distribution of ineffectiveness.

Nonparametric estimation of the conditional distribution function $F(Y|Z=z)$ requires the specification of weight functions and bandwidths. Kernel weight functions are used to give more weight to observations near the observation point. Bandwidths impose the window of localization. Literature shows that the choice of weighting function is far less important than the choice of the bandwidth - which we will discuss below.

We use kernel weights (l^c, l^u, l^o) with bandwidths (h^c, h^u, h^o) to specify the weight function for $z = [z^c, z^u, z^o]$, where z^c is a vector of continuous values, z^u is a vector of unordered discrete values, z^o is a vector of ordered discrete values. In specific, we specify an epanechnikov kernel function l^c to weight the continuous variables z^c (see (5.1)). An Aitchison and Aitken (1976) kernel l^u is specified to weight discrete unordered variables z_l^u with c_l categories (see (5.2)). In the extreme case of $h_l^u = 0$, no weight is given to observations with a different value of Z. The other extreme of $h_l^u = (c_l - 1)/c_l$ means that observations with $Z_{jl} \neq z_l$ receive equal weight as observations with $Z_{jl} = z_l$. In other words, z_l^u is ignored by the model. To weight the ordered discrete values z^o, we use a Wang and van Ryzin (1981) kernel function (see (5.3)).

[24] It is well known that in an input-oriented FDH approach with variable returns to scale (VRS), observation c is benchmarked against observations with $Y \geq y_c$. In other words, only the subsample with $Y \geq y_c$ determines the performance score of police department c. Our base 'BoD' model can be formulated as an input-oriented constant returns to scale (CRS) DEA model with a 'dummy input' always equal to 1 (see Cherchye et al. (2007)). It can easily be shown that in a CRS DEA model, also observations with $Y < y_c$ can influence the effectiveness score of observation c. Therefore, in contrast to the VRS-based order-m approach of Cazals et al. (2002), we draw observations from the whole sample and not only from the subsample of observations with $Y \geq y_c$.

$$l^c\left(\frac{Z^c_{jp}-z^c_p}{h^c_p}\right)=\begin{cases}\frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}\left(\frac{Z^c_{jp}-z^c_p}{h^c_p}\right)^2\right) & \text{if } \left(\frac{Z^c_{jp}-z^c_p}{h^c_p}\right)^2\leq 5\\ 0 & \text{otherwise.}\end{cases} \tag{5.1}$$

$$l^u(Z^u_{jl},z^u_l,h^u_l)=\begin{cases}1-h^u_l & \text{if } Z^u_{jl}=z^u_l,\\ h^u_l/(c_l-1) & \text{otherwise.}\end{cases} \tag{5.2}$$

$$l^o(Z^o_{ir},z^o_r,h^o_r)=\begin{cases}1 & \text{if } Z^o_{jr}=z^o_r,\\ (h^o_r)^{|Z^o_{jr}-z^o_r|} & \text{otherwise.}\end{cases} \tag{5.3}$$

To allow for a multivariate estimation, we use - as is common practice - product kernels. The product kernel of z^c is $W_{h^c}(Z^c_j,z^c)=\prod_{p=1}^{q}(h^c_p)^{-1}l^c((Z^c_{jp}-z^c_p)/h^c_p)$. For z^u, the product kernel is defined as $L_{h^u}(Z^u_j,z^u)=\prod_{l=1}^{v}l^u(Z^u_{jl},z^u_l,h^u_l)$. The product kernel of z^o is $L_{h^o}(Z^o_j,z^o)=\prod_{r=1}^{s}l^o(Z^o_{jr},z^o_r,h^o_r)$. All together, we can specify a Racine and Li (2004) generalized kernel function as $\mathcal{K}_h(Z_j,z)=W_{h^c}(Z^c_j,z^c)L_{h^u}(Z^u_j,z^u)L_{h^o}(Z^o_j,z^o)$, with $h=(h^c,h^u,h^o)$.

We estimate the conditional CDF following Li and Racine (2007,p. 184) by smoothing in direction of both Y as Z (see (5.4)). The optimal level of bandwidth is chosen by minimization of the integrated squared error (i.e., leave-one-out Least-Squares Cross-Validation).

$$\hat{F}(y|z)=\frac{n^{-1}\sum_{j=1}^{n}W_{h_y}\left(\frac{Y_j-y}{h_y}\right)\mathcal{K}_h(Z_j,z)}{n^{-1}\sum_{j=1}^{n}\mathcal{K}_h(Z_j,z)}. \tag{5.4}$$

Using the optimal bandwidth vector h, we construct $\hat{E}_m(y|z)$ by performing the following iteration process[25]:

[1] Draw for a given police department c, a sample of size m with replacement and with a probability $\mathcal{K}_h(Z_j,z)$. Denote this sample by $\Upsilon^{m,z,b}_c=\left\{Y^{z,b}_1,...,Y^{z,b}_m\right\}$.

[25] Analogously to Daraio and Simar (2007a)

[2] Solve the linear program:

$$\tilde{E}_m^{z,b}(y) = \max_{w_{c,1},\ldots,w_{c,q}} \sum_{i=1}^{q} w_{c,i} y_{c,i}^b \qquad s.t. \tag{B.5}$$

$$\sum_{i=1}^{q} w_{c,i} y_{j,i}^b \leq 1 \qquad \forall j = 1,\ldots,x,\ldots,m \tag{B.5a}$$

$$w_{c,i} \geq 0 \qquad \forall i = 1,\ldots,q \tag{B.5b}$$

$$\alpha_i \leq \frac{w_{c,i} y_{c,i}^b}{\sum_{i=1}^{q} w_{c,i} y_{c,i}^b} \leq \beta_i \qquad \forall i = 1,\ldots,q. \tag{B.5c}$$

[3] Redo [1] and [2] for b=1,...,B.

[4] Construct $\hat{E}_m(y|z) \approx \frac{1}{B}\sum_{b=1}^{B} \tilde{E}_m^{z,b}(y)$.

Inference on the impact of Z

To visualize the effect of Z on the production process, we use a nonparametric local-linear regression of Z on $\hat{Q}(y) = \hat{E}_m(y|z)/\hat{E}_m(y)$ as proposed by Daraio and Simar (2005) and Daraio and Simar (2007b). This can be formulated as a localized least squares regression:

$$\min_{\{a,b\}} \sum_{j=1}^{n} (\hat{Q}_j - a - (Z_j - z)'b)^2 \mathcal{K}_{\lambda}(Z_j, z), \tag{B.6}$$

with $\mathcal{K}_{\lambda}(Z_j, z)$ the generalized Li-Racine kernel weight function with bandwidth $\lambda = [\lambda^c, \lambda^u, \lambda^o]$. For the evaluation points $\{z_1,\ldots,z_k,\ldots,z_K\}$, we estimate the fitted values $\hat{\pi}^{z_k} = E[\hat{Q}_m|Z = z_k]$.[26] A $\hat{\pi}_m^{z_k}$ that increases (decreases) with Z, holding everything else equal, indicates that the environmental variable has a negative (positive) effect on effectiveness.[27]

Bootstrapping is used to construct confidence regions on the estimated fitted values $\hat{\pi}$. As we do not observe $Q(y|z)$, but only the estimate $\hat{Q}(y|z)$, the i.i.d. assumption is invalid and standard bootstrap theory cannot be applied on the pairs $(Z_j, \hat{Q}(Y_j|Z_j))$ (Badin et al., 2010b). We use the

[26] We also estimate the observation-specific coefficients $\hat{b}$, these are available upon request.

[27] We make use of the R package 'np' of Hayfield and Racine (2008) to estimate the CDF and local-linear regression.

Badin et al. (2010b) subsample bootstrap approach, which draws BB times $M < n$ observations directly from the observed and i.i.d. sample (Y_j, Z_j). For each sample bb of size M, the fitted values $\hat{\pi}^{*,bb,z_k}$, for $k = 1,...,K$ are estimated. Quantiles $(q^{*,z_k}_{M;\alpha/2}, q^{*,z_k}_{M;1-\alpha/2})$ of $\hat{\pi}^{z_k}_n - \hat{\pi}^{*,bb,z_k}_M$ are rescaled to correct for the difference in sample size and determine the $1-\alpha$ confidence interval of π^{z_k} :

$$\pi^{z_k}(\Upsilon) \in \left[\hat{\pi}^{z_k}_n - q^{*,z_k}_{M;\alpha/2}, \hat{\pi}^{z_k}_n - q^{*,z_k}_{M;1-\alpha/2})\right]. \tag{B.7}$$

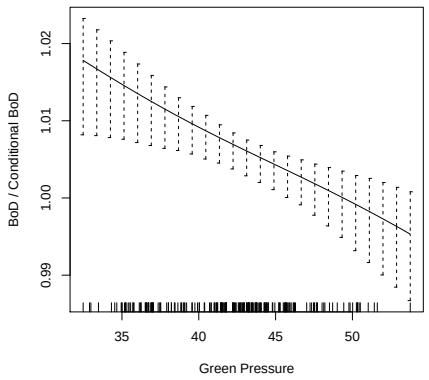
It is important to note that we have a subsample bootstrap of B replications (to construct robust conditional effectiveness scores) in a subsample bootstrap of BB replications (to allow for inference). Logically, computational burden increases dramatically with the size of B and BB. Preliminary analysis showed that setting B=100 and BB=200 suffices for robust inference in our setting. M is set to 150. Results are not sensitive for altering the value of M to 200 and 175.

5.6.3 Sensitivity tests

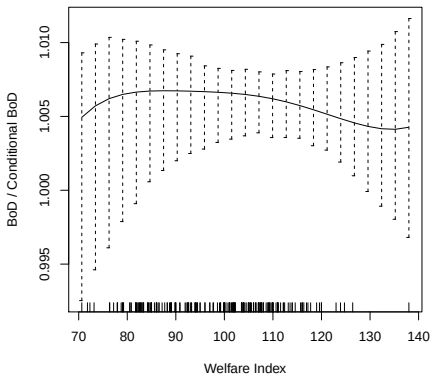
Altering the model specification

	Average	St.Dev	Min	Q1	Median	Q3	Max
Conditional BoD (Model 2)	0.932	0.044	0.799	0.899	0.938	0.967	1.001
Conditional BoD (Model 3)	0.933	0.045	0.798	0.901	0.942	0.967	1.004

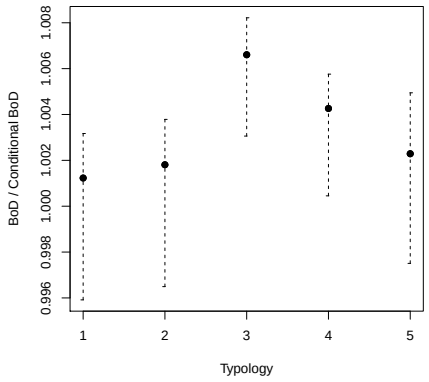
Table 5.4: Estimates of local police effectiveness in different model specifications (n=209 local police forces)



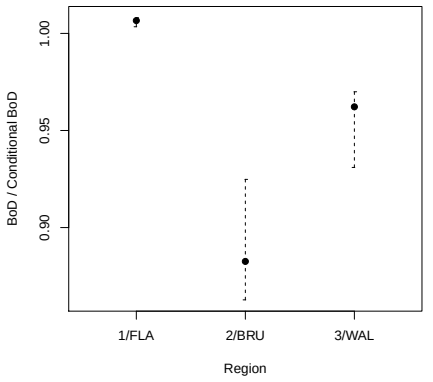
(a) Green Pressure



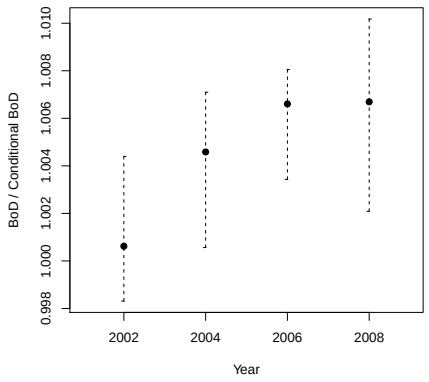
(b) Welfare Index



(c) Typology

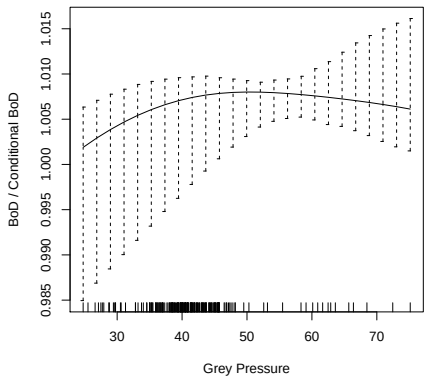


(d) Region

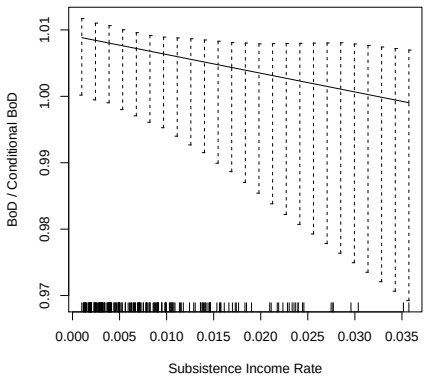


(e) Year

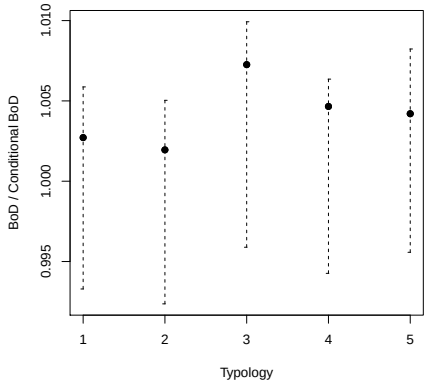
Figure 5.7: Visualization of the results (Model 2)



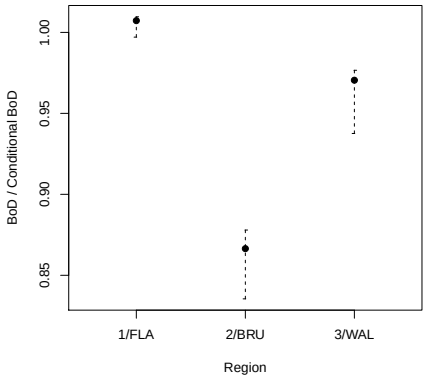
(a) Grey Pressure



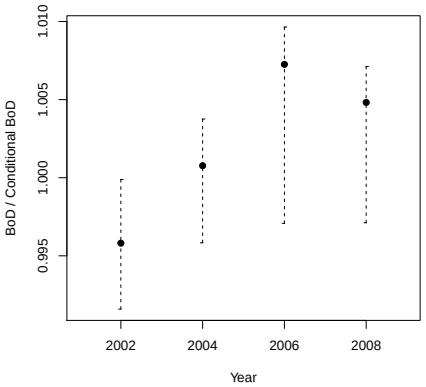
(b) Subsistence Income Rate



(c) Typology



(d) Region



(e) Year

Figure 5.8: Visualization of the results (Model 3)

Effectiveness analysis with dependent data

As our dataset consists of 209 observations from 84 police departments, we have dependent data. The made i.i.d. assumption is thus unrealistic in this setting. However, our results still hold when we relax the independence assumption in the used routines.

Li and Racine (2007, Chapter 18) review a large literature showing that consistency, the rates of convergence and asymptotic normality of nonparametric density, regression and conditionally density estimators, made under an i.i.d. assumption, still hold when we allow for so called 'weakly dependent' data. In other words, if we assume that the dependency disappears between $E_{c,t}$ and $E_{c,t+\tau}$ if τ goes to infinity, the used nonparametric inference is still valid. In particular, the leading bias and variance terms of the least squares cross-validation routine to settle the bias-variance trade-off are the same as in an i.i.d. setting, which implies the used cross-validation routine is asymptotically consistent and still valid if we relax the independence assumption.

Further, our subsample bootstrap routine relies on the assumption we draw i.i.d. observations. To test sensitivity for relaxing the independence assumption, we discuss the number of peer units in a bootstrap replication and we alter our bootstrap routine accordingly.
First, the localized subsample bootstrap routine could result in a situation where unit c from year l is only benchmarked against observations from unit c. Differently put, if the bandwidths are very small, it is possible that epanechnikov kernel weighting (with compact support) implies no probability to be drawn for observations from other units and positive probability to be drawn for observations from the unit in question. To show this is unlikely in our setting, we show in Table D.2 an overview of respectively the number of different observations and different units to be drawn in each of the 500 replications of size m for an observation. As observations from different regions are estimated to be non-comparable, the number of peer units is very low for departments in Brussels (as there are only 6 departments sampled). But still, no observation is ever compared solely to itself or other observations from the same unit. For units from the Walloon region and Flanders, the minimum of peer units from all replications for an observation is respectively 8 and 10. Half of the observations have no replication with less than respectively

14 and 17 peer units. The average of the median number of peer units over the 500 replications for an observation is respectively 19.46 and 23.67.

Second, we test sensitivity of the benchmarking routine for the presence of other observations from the same unit by altering the subsample bootstrap routine. In particular, we do not allow that in the evaluation of unit c from year l, observations from unit c from a different year are drawn. Differently put, unit c from year l is only benchmarked against other units and itself, but not against other observations from unit c. As such, we have a bootstrap estimate that controls the evaluation of unit c for possible impact of multiple observations from unit c. The conditional efficiency estimates and estimates of the effect of environmental variables are not sensitive for changing the bootstrap routine in this manner. The correlation with the base model is 0.997.

Although we showed empirical indications that our results are not sensitive for relaxing the independence assumption, the use of a bootstrap routine that allows for dependency in the data structure is still advised. However, to our knowledge, there exists no widely accepted bootstrap routine that allows for dependency and is directly applicable in this setting. Further research is needed to test the appropriateness of for example a 'block' bootstrap routine (used in time series studies to solve the temporal decency problem)[28] in a benchmarking setting.

[28]See Lahiri (2003) for an overview of resampling approaches that relax the i.i.d. assumption.

	Mean	St.Dev.	Min.	25%	Med.	75%	Max.
Peer observations							
Minimum of 500 replications per observation							
Flanders	24.69	2.12	19.00	23.00	25.00	26.00	29.00
Walloon region	19.72	1.91	14.00	19.00	20.00	21.00	23.00
Brussels	9.20	1.24	7.00	8.75	9.00	10.00	11.00
Median							
Flanders	33.02	1.76	27.00	32.00	33.75	34.00	35.00
Walloon region	27.51	1.91	21.00	27.00	28.00	28.25	30.00
Brussels	13.65	1.27	11.00	13.00	14.00	14.25	16.00
Maximum							
Flanders	41.08	1.84	34.00	40.00	41.00	42.00	46.00
Walloon region	35.46	2.03	28.00	35.00	36.00	36.50	40.00
Brussels	18.05	1.19	16.00	17.00	18.50	19.00	20.00
Peer units							
Minimum							
Flanders	16.97	2.07	10.00	16.00	17.00	18.00	21.00
Walloon region	13.72	1.68	8.00	13.00	14.00	15.00	16.00
Brussels	4.00	0.73	3.00	3.75	4.00	4.25	5.00
Median							
Flanders	23.67	2.00	18.00	22.00	24.00	25.00	27.00
Walloon region	19.46	1.69	13.00	19.00	20.00	20.25	23.00
Brussels	5.50	0.51	5.00	5.00	5.50	6.00	6.00
Maximum							
Flanders	30.38	2.25	24.00	29.00	31.00	32.00	35.00
Walloon region	25.24	1.82	19.00	24.00	25.00	26.50	29.00
Brussels	6.00	0.00	6.00	6.00	6.00	6.00	6.00

Table 5.5: Sensitivity test for relaxing independence assumption

5.6.4 Estimated optimal bandwidth sizes

	Socioeconomic char.	Demographic char.	Typology	Year	Region
Conditional BoD					
Base Model	$1.021e^{4}$	5.520	0.702	0.586	$7.864e^{-16}$
Model 2	17.210	5.428	0.551	0.536	$5.860e^{-16}$
Model 3	0.017	10.554	0.552	0.503	$1.700e^{-15}$
Nonparametric regression					
Base Model	$2.047e^{3}$	$2.265e^{6}$	0.548	0.484	$1.339e^{-15}$
Model 2	13.112	6.382	0.133	0.390	0.002
Model 3	$1.238e^{3}$	12.206	0.275	0.449	$2.327e^{-14}$

Table 5.6: Estimated optimal bandwidth sizes

References

Aitchison, J., Aitken, C. G. G., 1976. Multivariate binary discrimination by kernel method. Biometrika 63 (3), 413–420.

Badin, L., Daraio, C., Simar, L., 2010a. How to measure the impact of environmental factors in a nonparametric production model? ISBA Discussion Paper Series (1050).

Badin, L., Daraio, C., Simar, L., 2010b. Optimal bandwidth selection for conditional efficiency measures: A data-driven approach. European Journal of Operational Research 201 (2), 633–640.

Beck, K., Boni, N., Packer, J., 1999. The use of public attitude surveys: What can they tell police managers? Policing: An International Journal of Police Strategies & Management 22 (2), 191–213.

Brandl, S., Frank, J., Wooldredge, J., Watkins, R., 1997. On the measurement of public support for the police: A research note. Policing: An International Journal of Police Strategies & Management 20 (3), 473–480.

Bridenball, B., Jesilow, P., 2008. What matters the formation of attitudes toward the police. Police Quarterly 11 (2), 151–181.

Cao, L., Frank, J., Cullen, F., 1996. Race, community context and confidence in the police. American Journal of Police 15, 3–22.

Cao, L. Q., Stack, S., Sun, Y., 1998. Public attitudes toward the police: A comparative study between Japan and America. Journal of Criminal Justice 26 (4), 279–289.

Carter, D., 2002. The Police and the Community, 7th Edition. Englewood Cliffs, NJ: Prentice Hall.

Cazals, C., Florens, J. P., Simar, L., 2002. Nonparametric frontier estimation: A robust approach. Journal of Econometrics 106 (1), 1–25.

Charnes, A., Cooper, W. W., Rhodes, E., 1978. Measuring efficiency of Decision-Making Units. European Journal of Operational Research 2 (6), 429–444.

Charnes, A., Cooper, W. W., Rhodes, E., 1981. Evaluating program and managerial efficiency - An application of Data Envelopment Analysis to program follow through. Management Science 27 (6), 668–697.

Cherchye, L., De Borger, B., Van Puyenbroeck, T., 2006. Nonparametric tests of optimizing behavior in public service provision: Methodology and an application to local public safety. Tijdschrift voor Economie en Management 51 (3), 281–308.

Cherchye, L., Lovell, C. A. K., Moesen, W., Van Puyenbroeck, T., 2007. One market, one number? A composite indicator assessment of EU internal market dynamics. European Economic Review 51 (3), 749–779.

Daouia, A., Gijbels, I., 2011. Robustness and inference in nonparametric partial frontier modeling. Journal of Econometrics 161 (2), 147–165.

Daouia, A., Ruiz-Gazen, A., 2006. Robust nonparametric frontier estimators: Qualitative robustness and influence function. Statistica Sinica 16 (4), 1233–1253.

Daraio, C., Simar, L., 2005. Introducing environmental variables in nonparametric frontier models: A probabilistic approach. Journal of Productivity Analysis 24 (1), 93–121.

Daraio, C., Simar, L., 2006. A robust nonparametric approach to evaluate and explain the performance of mutual funds. European Journal of Operational Research 175 (1), 516–542.

Daraio, C., Simar, L., 2007a. Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications. Studies in Productivity and Efficiency. Springer.

Daraio, C., Simar, L., 2007b. Conditional nonparametric frontier models for convex and nonconvex technologies: A unifying approach. Journal of Productivity Analysis 28 (1-2), 13–32.

De Witte, K., Rogge, N., 2010. Dropout from secondary education: All's well that begins well. European Journal of Education - in Press.

Drake, L., Simper, R., 2002. X-efficiency and scale economies in policing: a comparative study using the distribution free approach and DEA. Applied Economics 34 (15), 1859–1870.

Drake, L., Simper, R., 2003. The measurement of English and Welsh police force efficiency: A comparison of distance function models. European Journal of Operational Research 147 (1), 165–186.

Drake, L., Simper, R., 2004. The economics of managerialism and the drive for efficiency in policing. Managerial and Decision Economics 25, 509–523.

Drake, L., Simper, R., 2005. The measurement of police force efficiency: An analysis of UK home office policy. Contemporary Economic Policy 23, 465–482.

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

Gaines, L. K., Kappeler, V. E., Vanghn, J. B., 1997. Policing in America, 2nd Edition. Cincinnati, OH: Anderson Publishing Company.

Hayfield, T., Racine, J. S., 2008. Nonparametric econometrics: The np package. Journal of Statistical Software 27 (5), 1–32.

Hesketh, B., 1992. The police use of surveys: Valuable tools or misused distractions? Police Studies 15, 55–61.

Hwang, E. G., McGarrell, E. F., Benson, B. L., 2005. Public satisfaction with the South Korean Police: The effect of residential location in a rapidly industrializing nation. Journal of Criminal Justice 33 (6), 585–599.

Jeong, S., Park, B., Simar, L., 2010. Nonparametric conditional efficiency measures: Asymptotic properties. Annals of Operations Research 173 (1), 105–122.

Jesilow, P., Meyer, J., 2001. The effect of police misconduct on public attitudes: A quasi-experiment. Journal of Crime and Justice 24 (1), 109–121.

Kusow, A., Wilson, L., Martin, D., 1997. Determinants of citizen satisfaction with the police: The effects of residential location. Policing: An International Journal of Police Strategy and Management 20, 655–664.

Lahiri, S., 2003. Resampling Methods for Dependent Data. Springer Series in Statistics. Springer-Verlag, New York.

Li, Q., Racine, J., 2007. Nonparametric Econometrics: Theory and practice. Princeton University Press.

Lovell, C., Pastor, J., 1999. Radial DEA models without inputs or without outputs. European Journal of Operational Research 118 (1), 46–51.

Marenin, O., 1983. Supporting the local police: The differential group basis of varieties of support. Police Studies 6, 50–56.

Melyn, W., Moesen, W., 1991. Towards a synthetic indicator of macroeconomic performance: Unequal weighting when limited information is available. Tech. Rep. 17, Public Economics Research paper, CES, KU Leuven.

Murphy, D. W., Worrall, J. L., 1999. Residency requirements and public perceptions of the police in large municipalities. Policing: An International Journal of Police Strategies & Management 22 (3), 327–342.

Murty, K. S., Roebuck, J. B., Smith, J. D., 1990. The image of the police in black atlanta communities. Journal of Police Science and Administration 17 (4), 250–257.

Nicholson-Crotty, S., O'Toole, L. J., 2004. Public management and organizational performance: The case of law enforcement agencies. Journal of Public Administration Research and Theory 14 (1), 1–18.

Nyhan, R., Martin, L., 1999. Assessing the performance of municipal police service using data envelopment analysis: An exploratory study. State and Local Government Review 31 (1), 18–30.

Portela, M., Thanassoulis, E., 2001. Decomposing school and school type inefficiencies. European Journal of Operational Research 132, 357–373.

Racine, J., Li, Q., 2004. Nonparametric estimation of regression functions with both categorical and continuous data. Journal of Econometrics 119 (1), 99–130.

Reisig, M., Correia, M., 1997. Public evaluations of police performance: An analysis across three levels of policing. Policing: An International Journal of Police Strategies and Management 20, 311–325.

Reisig, M. D., Parks, R. B., 2000. Experience, quality of life, and neighborhood context: A hierarchical analysis of satisfaction with police. Justice Quarterly 17 (3), 607–630.

Rogge, N., Verschelde, M., 2012. A composite index of citizen satisfaction with local police services. HUB Research paper.

Scaglion, R., Condon, R. G., 1980. Determinants of attitudes toward city police. Criminology 17 (4), 485–494.

Schafer, J., Huebner, B., Bynum, T., 2003. Citizen perceptions of police services: Race, neighbourhood context, and community policing. Policy Quarterly 6, 440–467.

Simar, L., Wilson, P., 2008. Statistical inference in nonparametric frontier models: recent developments and perspectives. In: Fried, H., Lovell, C., Schmidt, S. (Eds.), The measurement of productive efficiency. Oxford University Press.

Sims, B., Hooper, M., Peterson, S. A., 2002. Determinants of citizens' attitudes toward police - Results of the Harrisburg citizen survey - 1999. Policing: An International Journal of Police Strategies & Management 25 (3), 457–471.

Sullivan, P. S., Dunham, R. G., Alpert, G. P., 1987. Attitude structures of different ethnic and age-groups concerning police. Journal of Criminal Law & Criminology 78 (1), 177–196.

Sun, S., 2002. Measuring the relative efficiency of police precincts using Data Envelopment Analysis. Socioeconomic Planning Sciences 36, 51–71.

Thanassoulis, E., 1995. Assessing police forces in England and Wales using Data Envelopment analysis. European Journal of Operational Research 87 (3), 641–657.

Thanassoulis, E., Portela, M., Allen, R., 2004. Incorporating value judgements in DEA. In: Cooper, W., Seiford, L., Zhu, J. (Eds.), Handbook on Data Envelopment Analysis. Kluwer Academic Publishers, Dordrecht, pp. 99–138.

Van Den Bogaerde, E., Van Den Steen, I., De Bie, A., Klinckhamers, P., Vandendriessche, M., 2009. Veiligheidsmonitor 2008-2009: Analyse van de federale enquête. Tech. rep., Federale Politie, Directie Operationele Politionele Informatie, Politiebeleidsondersteuning, 57 p.

Van Den Bogaerde, E., Van Den Steen, I., Klinckhamers, P., Vandendriessche, M., 2007. Veiligheidsmonitor 2006: Analyse van de federale enquête. Tech. rep., Federale Politie, Algemene Directie Operationele Ondersteuning, Directie van de nationale gegevensbank, 56 p.

Wang, M. C., van Ryzin, J., 1981. A class of smooth estimators for discrete-distributions. Biometrika 68 (1), 301–309.

Webb, V., Katz, C., 1997. Citizen ratings of the importance of community policing activities. Policing: An International Journal of Police Strategy and Management 20, 7–23.

Webb, V., Marshall, C., 1995. The relative importance of race and ethnicity on citizen attitudes toward the police. American Journal of Police 14 (2), 45–66.

Worrall, J., 1999. Public perceptions of police efficacy and image: The "fuzziness" of support for the police. American Journal of Criminal Justice 24, 47–66.

Wu, T. H., Chen, M. S., Yeh, J. Y., 2010. Measuring the performance of police forces in Taiwan using Data Envelopment Analysis. Evaluation and Program Planning 33 (3), 246–254.

Zamble, E., Annesley, P., 1987. Some determinants of public-attitudes toward the police. Journal of Police Science and Administration 15 (4), 285–290.

Zevitz, R., Rettammel, R., 1990. Elderly attitudes about police services. American Journal of Police 9, 25–39.

6

Noise, Inefficiency, and Nonparametric Bank Branch Evaluation[1]

6.1 Introduction and related literature

To facilitate strategic decisions in times when banks in the developed world face growing doubts and many banks are directly or indirectly controlled by governments, both policy makers and bank managers need a detailed understanding of bank efficiency at the branch level. Both the banking sector and individual banks could considerably improve their efficiency by better understanding why some branches do better than others. In this paper we provide a novel approach to evaluate bank branch efficiency. Bank performance is measured in an innovative way, by also

[1]This chapter is the result of joint work with Koen Schoors and Paul Gemmel.

including measures of loyalty that look beyond the narrow approach of banks as selling outlets. We combine insights from the stochastic frontier (SFA) literature and the Data Envelopment Analysis (DEA) literature to arrive at a very robust identification of underperforming branches. More specifically we propose to use combined information from a deterministic conditional robust frontier approach and a nonparametric stochastic frontier approach to unequivocally identify underperforming branches, allowing for heterogeneity, nonlinearities, environmental variables and uncertain noise. We demonstrate our approach in a study of market efficiency of 717 bank branches of a large bank in Belgium.

Until recently ratio measures were broadly used to asses bank branches by virtue of their simplicity. But ratio measures implicitly assume constant returns-to-scale and are either one-dimensional or combine multiple dimensions into an unsatisfactory single number using arbitrary weights. The outcomes of ratio analysis may therefore be confusing or even contradictory and do not offer a clear assessment of bank branch performance (Paradi et al., 2004). Operational Research (O.R.) techniques (DEA[2] and SFA[3]) address the shortcomings of traditional ratio measures. These methods allow to estimate and compare the efficiency of multiple input -multiple output production technologies assuming variable returns to scale. By taking into account multiple inputs and outputs simultaneously, such multivariable models have the ability to detect the branches that produce more or better outcome using the same resources (inputs) and operate under the same conditions, or produce the same outcome using less resources. Berger and Humphrey (1997) review the use of O.R. approaches to examine financial institutions. Berger (2007) reviews international comparisons of bank efficiency. Fethi and Pasiouras (2010) review 196 studies that use O.R. or Artificial Intelligence techniques to assess bank performance, whereof 151 use DEA and 30 are at branch level.

Bank efficiency requires efficient intermediation between capital buyers and sellers (intermediation or market efficiency) and the production of bank services at minimum cost (production or cost efficiency). One of the main differences between these two production function approaches

[2]Initiated by Farrell (1957) and operationalized as linear programming estimators by Charnes et al. (1978).

[3]Meeusen and Van Den Broeck, 1977 and Aigner et al., 1977.

is that intermediation efficiency treats deposits as an input, while production efficiency treats deposits as an output (see Berger and Humphrey (1997) for an early overview). Bank branches play a crucial role in production efficiency, because operational resources are still largely converted in client services and transactions at the bank branch level. Estimating production efficiency at the branch level can therefore greatly improve our understanding of bank level efficiency. A fair evaluation of bank branch production efficiency requires (1) that bank branch managers are only held accountable for the stated objective of the bank branch, (2) that managers are only accountable for what is discretionary and (3) that the estimation approach minimizes the imposition of arbitrary assumptions.

6.1.1 Bank branch objectives

Our understanding of the proper bank branch objectives has evolved over time. The early literature viewed bank branches as 'convenience outlets' (see e.g. the seminal work of Berger et al. (1997)). In this view bank branches naturally focus on the minimization of costs to provide convenience to bank customers. Hirtle (2007) notes how this delivery model was challenged by technical (e.g. Internet, ATM's) and regulatory innovations in the 1990s. As a result of these pressures bank branches evolved from predominantly transaction-based entities to more sales-oriented entities (Portela and Thanassoulis, 2007). The recent literature fully acknowledges he crucial role of bank branches as 'selling outlets' (see e.g. Athanassopoulos (1998), Cook et al. (2000), Cook and Hababou (2001) and Portela and Thanassoulis (2007)). Athanassopoulos (1998) states that "*the primary objective of a bank branch is to penetrate its market by selling financial products to new costumers while delivering services to existing customers*".

To ensure long term revenue generation however, establishing a relation with your bank customer and building bank loyalty may be more important than a narrow focus on immediate sales targets. Consumer retention has indeed become much more troublesome because of technological innovations, such as internet and online banking, allow customers to evaluate competing alternatives at a touch of a button and to act immediately upon such information (Camanho and Dyson, 2005). Banks address this problem by investing in intangible and bank-customer specific assets

to establish interpersonal relations with their client and stimulate their loyalty to the bank, which in turn improves the long term revenue generation of the bank. Bank branches play a crucial role in building and maintaining customer loyalty. This approach to banking is usually referred to as relationship banking.[4] Degryse and Ongena (2001) show that borrowers with multiple bank relationships are less profitable than those that borrow predominantly from one bank. Taking into account all three aspects, we can define the objective of the bank as "*to penetrate its market by selling financial products to new costumers, while tying profitable costumers to the bank and delivering services to existing customers*". We therefore follow Portela and Thanassoulis (2007) and include client loyalty directly as a bank branch output.

6.1.2 Measuring inefficiency

As mentioned higher, methods from two methodological families are applied in the estimation of bank branch efficiency. The first family of methods is based on non-parametric Data Envelopment Analysis (DEA). The vast majority of OR studies on bank branch efficiency uses a DEA-based approach. In standard DEA the data fully determine the shape of the frontier without any room for noise in the data. A bank branch is considered efficient if no other bank branch or convex combination of bank branches produces more output with the same or less operational costs. But this standard DEA analysis has become subject to several criticisms.

First, in the early OR literature on bank branch efficiency some studies found full efficiency for almost all branches solely because there were too many inputs, outputs and environmental variables relative to the sample size (Berger et al., 1997). This dimensionality problem is adequately addressed by using larger sample sizes, as will be the case in this study.

Second, traditional DEA models did not allow to control for environmental variables. This makes it problematic to use these models for real life bank branch benchmarking because it is inappropriate to hold bank branch managers accountable for factors beyond their span of control. We

[4]For an overview of the benefits of relationship banking for the bank see Sharpe (1990), Rajan (1992), Petersen and Rajan (1995), Ongena and Smith (2001), Rajan and Zingales (2003), Elyasiani and Goldberg (2004), Freixas (2005), Degryse and Ongena (2005).

control for environmental variables by also applying conditional efficiency approaches that use kernel weights to ensure that bank branches are benchmarked against peers with similar environments.

Third, DEA estimates are very sensitive to extreme observations as the frontier envelops all data points. It is not fair to benchmark bank branches against one or more outlying observations that may be driven by measurement error or unobserved environmental differences. The recent nonparametric partial frontier approaches proposed by Cazals et al. (2002), Aragon et al. (2005) and Daouia and Simar (2007) address this issue of a more robust frontier by not enveloping all data points. The basic idea of these methods is to estimate a partial frontier - close to the full frontier - that gives extreme observations less impact on the frontier estimates. Cazals et al. (2002) propose a robustified frontier by employing a subsample bootstrap replication approach on the subset of observations with $\mathbf{X} \leq \mathbf{x}$ (i.e., the order-m frontier). Daouia and Simar (2007) propose to use the α quantile of observations with $\mathbf{X} \leq \mathbf{x}$ as benchmark (i.e., the order-α quantile frontier). The robustness of the partial frontier approaches is demonstrated theoretically by Daouia and Ruiz-Gazen (2006) and Daouia and Gijbels (2011). However, although partial frontier approaches successfully deal with some of the problems of traditional DEA models, they are still fully deterministic by nature and neglect the possibility of noise. In result, no smooth decomposition of noise and inefficiency can be made.

Parametric stochastic frontier approaches, the second methodological family in the estimation of bank branch efficiency, have been developed specifically to accommodate noise in the data generation process. To smoothly decompose noise from inefficiency, standard stochastic frontier analysis (SFA) however imposes (1) the functional form of the frontier (e.g., Cobb-Douglas, Translog, Fourier), (2) the distribution of noise and (3) the distribution of inefficiency (e.g., half-normal, truncated normal, exponential, gamma). A survey by Yatchew (1998) clearly indicates that economic theory almost never specifies a precise specification of the functional form of a production function. As such, imposing an arbitrary functional specification of the production frontier can result in erroneous inference, which in turn biases the estimates and makes the analysis intricate.

Kumbhakar et al. (2007) proposed an alternative approach to loosen simultaneously the *a priori* assumptions on (1) the specification of the frontier, (2) the distribution of inefficiency and (3) the distribution of noise. They propose to localize the parametric stochastic frontier model, based on the local maximum likelihood approach of Tibshirani and Hastie (1987) and Fan et al. (1996). The resulting 'local maximum likelihood approach to estimate the stochastic frontier' (LMLSF) localizes the specification of the global frontier. Additionally, the approach is robust for unknown heteroskedasticity in both noise and inefficiency. The LMLSF method makes the parameters of a parametric model dependent on the covariates via a process of localization. In result the marginal frontier impact of inputs can be estimated for each data point.[5]

A direct implication of localization is that the frontier can be non-monotone or non-concave. Monotone, multivariate and concave estimates of nonparametric stochastic frontier can easily be achieved as shown in Simar and Wilson (2011). They extend the LMLSF approach to the full multivariate model without imposing parametric assumptions on the production relationship by the use of polar coordinates as in Simar (2007). They propose a two-step approach where the cloud of data points is pre-whitened from noise by a nonparametric stochastic frontier in the first step and inefficiency is measured as a distance to the pre-whitened frontier in a second step. Free disposability or concavity are imposed by applying respectively Free Disposal Hull (FDH) or Data Envelopment Analysis (DEA) in the second step.

Park et al. (2010) have extended the nonparametric stochastic frontier approach to allow for categorical environmental variables. However, the LMLSF implies (1) remaining distributional assumptions on inefficiency and noise for the anchorage model and (2) a high computational burden. It remains an open question whether it is fair to benchmark bank branches on the basis of an arbitrary decomposition of noise and inefficiency, even if it is only imposed locally. It remains unclear whether (locally) imposing distributional assumptions on noise and inefficiency is less

[5]The value of the LMLSF approach is shown in recent applications. Kumbhakar et al. (2007) have used the LMLSF approach to analyze the cost function of a random sample of 500 U.S. commercial banks. Additionally, Kumbhakar and Tsionas (2008) have applied the approach to estimate stochastic cost frontier models for a sample of 3691 U.S. commercial banks, while Serra and Goodwin (2009) use the approach to compare efficiency ratings of organic and conventional arable crop farms in the Spanish region of Andalucía.

problematic than ignoring noise altogether.

It is clear that the drawbacks of the SFA models correspond to the benefits of the DEA model, and vice versa. The recent semiparametric and nonparametric advances in both methods sketched above try to combine merits of both SFA and DEA while simultaneously limiting or eliminating the drawbacks. A fair evaluation of bank branches requires us to acknowledge the absence of consensus on how to decompose noise and inefficiency. Our contribution is that we use combined information from a deterministic robust frontier approach and a nonparametric stochastic frontier approach to allow a full understanding of (1) the effectiveness of resources and (2) the efficiency levels while allowing for heterogeneity, nonlinearities, environmental variables and (uncertain) noise.

6.2 Data

We use confidential data from a large Belgian bank. Our dataset was used by an academic team to assess bank branch performance in 2001-2003. This data is of exceptional quality as the dataset is the result of a six-month discussion between the academic group and the bank management. A lot of time was spent in making data uniform for all branches, regardless of their origin. A six-month pilot study with 180 branches was performed before running DEA models. A working group consisting of bank staff and management, district managers and the academic team was then formed to conduct the whole study. This team ended up with data for 2002 that was comparable across the Belgian territory and that could be used in a benchmarking analysis. In line with the research team, we limit our analysis to branches with a staff of maximum 15 FTE.[6] In addition, we exclude branches with no ATM's and branches that are less then 20 hours per week open to ensure full comparability. In total, our sample consists of 717 comparable bank branches located in Belgium.

[6]This because we believe that branches with a staff of more than 15 FTE have other objectives (e.g. more intermediation) and different inputs.

As in Athanassopoulos (1998) our model includes labor and technology facilities at the branches' disposal as bank branch inputs. We measure labor input as the number of full time equivalent staff at the branch level. Technology facilities are proxied by the number of ATM's per branch. In our data window this is appropriate, because there were at that time considerable branch level differences in the number of ATM's and the penetration of self-banking. In contrast to most of the literature on bank branch performance we also use market potential as an input variable, as suggested by Athanassopoulos (1998) and more recently Camanho and Dyson (2005). Our dataset contains detailed knowledge of the market potential on the city/quarter level. A 'jar of 10,000,000 potential points' are distributed over all the communities/quarters in Belgium. Of the total of 10,000,000 points, 7,500,000 points are distributed to capture the potential in the retail market. 1,250,000 points capture the potential of professionals and another 1,250,000 points capture the potential of small enterprises. See Appendix A for a detailed overview of the construction of the confidential bank branch level index of market potential.

	Mean	St.Dev.	Min.	Q1.	Med.	Q3	Max.
Input							
Staff in FTE	7.173	3.095	1.900	4.700	6.560	9.260	15.000
Market Potential	7.049	3.407	0.361	4.544	6.475	8.774	20.792
Number of ATM's	4.344	2.531	1.000	2.000	4.000	6.000	16.000
Output							
Sales composite	3.470	1.741	0.732	2.172	3.137	4.598	9.844
New credit to retail clients	149.338	92.325	21.000	80.000	127.000	192.000	518.000
New credit to prof. and small ent.	84.342	48.935	6.000	48.000	78.000	111.000	325.000
New insurances	183.378	109.067	11.000	105.000	161.300	239.300	658.900
New accounts	774.550	404.541	133.100	471.600	703.800	1021.700	3323.200
Loyalty composite	3.628	1.735	0.747	2.290	3.362	4.644	9.731
Retail clients in 3 or 4 d.	257.232	141.045	36.000	147.000	233.000	344.000	760.000
Retail clients HP in 3 or 4 d.	241.529	137.294	21.000	138.000	218.000	309.000	725.000
Prof. and small ent. in 3or 4 d.	46.561	24.027	5.000	28.000	42.000	61.000	142.000
Prof. and small ent. HP in 3 or 4 d.	65.798	38.056	4.000	37.000	61.000	87.000	237.000
Output modulus	5.038	2.421	1.168	3.143	4.598	6.607	13.511
Output mix (amplitude)	0.814	0.092	0.507	0.757	0.818	0.873	1.144

Table 6.1: Summary table input-output

As discussed above, the output of a bank branch is multidimensional. We capture output by a composite indicator of sales and a composite indicator of client loyalty. Table 6.1 shows that we have four variables to proxy the total sales of the bank branch and another four variables to proxy client loyalty. 4 domains are defined by the bank to measure the scope of activities: '*Save and Invest*', '*Credit*', '*Insurance*' and '*Daily banking*'. Clients that are active in 3 or 4 domains are considered to be loyal. As wealthy clients have the potential to be highly profitable for the bank, they are grouped in separate 'High Potential' categories.

To aggregate the 4-dimensional vector of Sales (Loyalty), $\mathbf{S}$ ($\mathbf{L}$), we follow Simar and Daraio (2007, p.149). We first normalize the data by dividing each sub-item by its standard deviation. Note that the Farrell-Debreu inefficiency measures radial distance and the estimates are scale-

invariant. This means that normalization of the data does not affect the efficiency results in the analysis. We then minimize the sum of squares of the residuals over $\mathbf{b}$ ($\mathbf{c}$) to find the projection of the scaled output data matrix $\mathbf{S}$ ($\mathbf{L}$) on vector $\mathbf{b}$ ($\mathbf{c}$) that best represents the data matrix $\mathbf{S}$ ($\mathbf{L}$). To obtain the Sales (Loyalty) aggregate $\mathbf{F}_{Sales}$ ($\mathbf{F}_{Loyalty}$), we multiply the first eigenvector $\mathbf{b}$ ($\mathbf{c}$) of the matrix $\mathbf{S}'\mathbf{S}$ ($\mathbf{L}'\mathbf{L}$) with vector $\mathbf{S}$ ($\mathbf{L}$).

$$\mathbf{F}_{Sales} = \mathbf{S}\,\mathbf{b} = b_1\mathbf{S}_1 + b_2\mathbf{S}_2 + b_3\mathbf{S}_3 + b_4\mathbf{S}_4 \tag{6.1}$$

$$\mathbf{F}_{Loyalty} = \mathbf{L}\,\mathbf{c} = c_1\mathbf{L}_1 + c_2\mathbf{L}_2 + c_3\mathbf{L}_3 + c_4\mathbf{L}_4. \tag{6.2}$$

Respectively 76% and 75% of inertia is explained by respectively the sales and loyalty composite. The high correlation between the original variables and the aggregate factors is shown in Table 6.2 and Table 6.3.[7]

	Sales composite
Correlation	
New credit to retail clients	0.888
New credit to professionals and small enterprises	0.810
New Insurances	0.860
New accounts	0.921
Inertia explained	0.758

Table 6.2: Sales composite

[7] As in Daraio and Simar (2007), we do not formally test the appropriateness of aggregation as the correlations are high enough for this particular application.

	Client loyalty composite
Correlation	
Retail clients in 3 or 4 domains	0.879
Retail clients High potential (HP) in 3 or 4 domains	0.868
Professionals and small enterprises in 3 or 4 d.	0.908
Professionals and small enterprises HP in 3 or 4 d.	0.813
Inertia explained	0.753

Table 6.3: Client loyalty composite

We want to control for environmental variables carefully in order to make sure that bank branches are only evaluated on the basis of factors in their span of control. Failing to control for the environment may indeed confound true managerial inefficiency with pure environmental factors and hence bias the efficiency estimates (see Bos and Kool (2006) and Bos et al. (2009)). We include three vectors of environmental variables. The first environmental vector is the client profile, that is largely driven by regional demography and migration. Table 6.4 illustrates that the client base of the sampled bank branches consists predominantly of families and seniors. However, there is considerable variation in client profile. Table 6.5 shows that a K-means clustering analysis reveals basically three profiles, namely 1) seniors, 2) small firms and professionals and 3) youngsters and families with (young) children. Although for some bank branches, the proportion of professionals and small enterprises is higher, the importance of professionals and small enterprises in the client base is for all bank branches below 38 percent.[8].

[8]As there is no specialization, we assume that the weights that the bank headquarter gives to the components of the sales and loyalty composite do not differ between bank branches

	Mean	St.Dev.	Min.	Q1	Median	Q3	Max.
Youngsters	14.14	3.01	5.70	12.30	14.40	15.90	29.40
Families	36.26	4.97	17.90	33.50	36.80	39.40	57.80
Seniors	36.84	5.28	12.10	33.60	36.80	39.90	55.40
Self-employed	5.66	2.83	1.00	3.80	5.00	6.80	21.20
Liberal professions	1.29	0.73	0.00	0.80	1.10	1.50	6.10
Small enterprises	5.81	3.06	1.10	3.80	5.10	7.10	25.30

Table 6.4: Client profile - summary statistics

	Cluster 1	Cluster 2	Cluster 3
Center Means			
Youngsters	-0.091	-1.241	0.719
Families	-0.183	-1.245	0.814
Seniors	0.575	0.384	-0.775
Self-employed	-0.307	1.027	-0.208
Liberal professions	-0.122	1.038	-0.401
Small enterprises	-0.294	1.379	-0.400
Within SS	620.021	1066.510	787.230
Group size	288	144	285

Table 6.5: K-means clustering

Also the urban environment matters for bank branch evaluation. It seems wise not to compare a bank branch in the heart of a world city like Brussels to a bank branch in a local non-urbanized village. Therefore we control for the urbanization typology of the bank branch location. Table 6.6 shows how the bank branches in this study are distributed across location typology. Last, we also want to control for purely regional effects, other than those reflected already indirectly in client profile and typology. Table 6.7 shows how our bank branches are distributed across Belgian regions. We include environmental controls for the three clusters of client profiles, the location typology and the region in all estimated models. Since competition is very strongly correlated to

region and typology of the location, it is also indirectly controlled by our environmental variables.

Typology	Group size
Large city	103
Regional city	90
Well equipped small city	58
Decently equipped small city	14
Moderately equipped small city	43
Well equipped non-urban city	217
Moderately equipped non-urban city	176
Weakly equipped non-urban city	16

Table 6.6: Typology of city, following Gemeentekrediet, 1998, "Actualisering van de stedelijke hiërarchie in België"

Region	1	2	3	4	5	6	7	8	9	10
Group size	77	90	93	82	39	52	89	84	58	53

Table 6.7: Region

This yields the model of bank branch production sketched in Figure 6.1, with three inputs (staff, physical capital and market potential), two composite indicators of output (sales and client loyalty) and three classes of environmental variables (client profile, urbanization typology and region-specific effects).

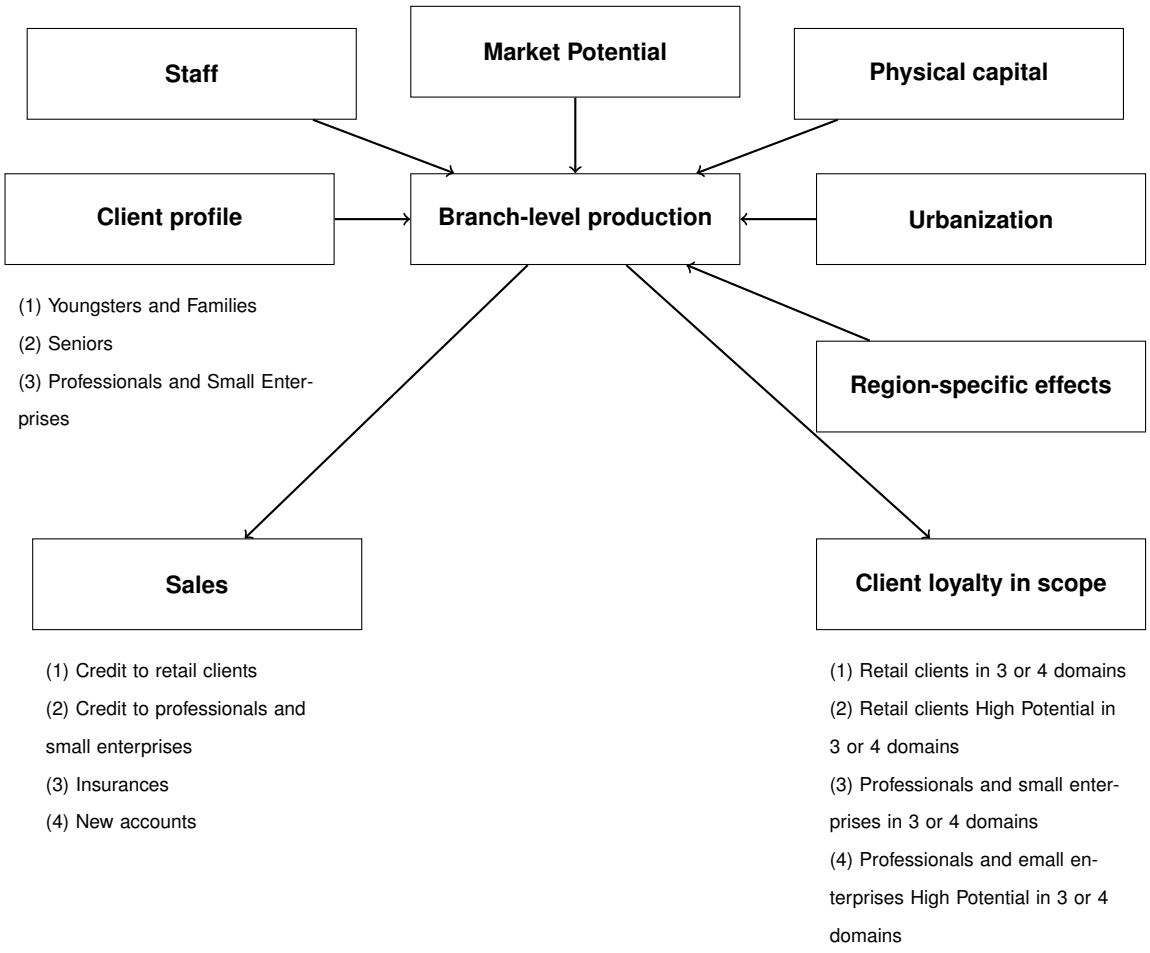


Figure 6.1: Branch-level production

6.3 Methodology

Since our approach relies on the combination of information from two methodological families, namely the deterministic conditional frontier and the multivariate stochastic nonparametric conditional frontier, we start by reviewing these two methodologies in more detail.

6.3.1 Deterministic conditional frontier

In this section, we discuss the output-orientated order-m frontier approach.[9] Assume that producers use a heterogeneous non-negative input vector $\mathbf{x} \in \Re_+^p$ to produce a heterogeneous output vector $\mathbf{y} \in \Re_+^q$. The production set Ψ of feasible input-output combinations can be defined as:

$$\Psi = \left\{ (\mathbf{x}, \mathbf{y}) \in \Re_+^{p+q} \mid \mathbf{x} \text{ can produce } \mathbf{y} \right\}. \tag{6.3}$$

The traditional 'Data Envelopment Analysis' (DEA; Charnes et al., 1978) literature estimates the production set while including all observed input-output combinations. As such, it estimates the efficiency of observations relatively to a full frontier. Farrell (1957) and Debreu (1951) were the first to acknowledge that the output-efficiency score (i.e., maximization of output $\mathbf{y}$ given the observed inputs $\mathbf{x}$) of an observation $(\mathbf{x}, \mathbf{y})$ can be obtained as:

$$\lambda(\mathbf{x}, \mathbf{y}) = \sup\{\lambda | (\mathbf{x}, \lambda\mathbf{y}) \in \Psi\}. \tag{6.4}$$

A value $\lambda(\mathbf{x}, \mathbf{y}) = 1$ indicates full technical efficiency (i.e., there are no observations which are able to produce more outputs for the given input set). A $\lambda(\mathbf{x}, \mathbf{y}) > 1$ indicates inefficiency, i.e., it is possible to have a radial increase of $\lambda(\mathbf{x}, \mathbf{y})$ in all the outputs in order to reach the efficient frontier.

Under the assumption of free disposability[10], probability theory can be used to interpret the efficiency scores. In particular, efficiency can be viewed as the proportional augmentation of output that unit $(\mathbf{x}, \mathbf{y}) \in \Psi$ needs to obtain in order to have a zero percent probability to be dominated, given the inputs $\mathbf{x}$. Following Cazals *et al.* (2002), this can be algebraically expressed as:

$$\lambda(\mathbf{x}, \mathbf{y}) = sup\{\lambda | S_{\mathbf{Y}|\mathbf{X}}(\lambda\mathbf{y}|\mathbf{x}) > 0\}, \text{ with } S_{\mathbf{Y}|\mathbf{X}} = \text{Prob}(\mathbf{Y} \geq \mathbf{y} | \mathbf{X} \leq \mathbf{x}). \tag{6.5}$$

By replacing in (6.5) the conditional survival function $S_{\mathbf{Y}|\mathbf{X}}$ by its empirical version $\hat{S}_{\mathbf{Y}|\mathbf{X}}$, Free Disposal Hull (FDH) inefficiency estimates $\hat{\lambda}_{FDH}(\mathbf{x}, \mathbf{y})$, as introduced in Deprins et al. (1984),

[9] Although the outline is limited to the output-oriented case, the extension to input-orientation or hyperbolic orientation is rather straightforward.

[10] i.e., if $(\mathbf{x}, \mathbf{y}) \in \Psi$, then any $(\mathbf{x}', \mathbf{y}')$ such that $\mathbf{x}' \geq \mathbf{x}$ and $\mathbf{y}' \leq \mathbf{y}$ is also in Ψ.

are obtained. If, additionally to FDH, a convexity assumption is imposed on the attainable set, one obtains the Data Envelopment Analysis (DEA) inefficiency estimates $\hat{\lambda}_{DEA}(\mathbf{x},\mathbf{y})$.

By definition, a DEA or FDH frontier envelops all n observations. Consequently, estimates are very sensitive to extreme data points and outliers. To address this problem, Cazals et al. (2002) propose the order-m frontier, which is defined as the expected frontier when considering a random set Ψ_m of only $m < n$ random observations with $\mathbf{X} \leq \mathbf{x}$. As atypical observations are not part of the sub sample Ψ_m in every draw, the impact of such observations on the inefficiency score $\lambda_m(\mathbf{x},\mathbf{y})$ is mitigated. If a bank branch is expected to perform better than m randomly drawn bank branches, it obtains a super-efficient value of $\lambda_m(\mathbf{x},\mathbf{y}) < 1$, otherwise $\lambda_m(\mathbf{x},\mathbf{y}) \geq 1$. Cazals et al. (2002) made clear that order-m inefficiency $\lambda_m(\mathbf{x},\mathbf{y})$ can be defined as a simple univariate integral function which only depends on the conditional survivor function $S_{\mathbf{Y}|\mathbf{X}}$ (see (6.6)). By replacing $S_{\mathbf{Y}|\mathbf{X}}$ by $\hat{S}_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ in (6.6), we obtain the order-m inefficiency estimate $\hat{\lambda}_m(\mathbf{x},\mathbf{y})$.

$$\lambda_m(\mathbf{x},\mathbf{y}) = \int_0^\infty \left[1-(1-S_{\mathbf{Y}|\mathbf{X}}(u\mathbf{y}|\mathbf{x}))^m\right] du. \tag{6.6}$$

As the performance of a bank branch depends among others on exogenous locational influences (see the review of Fethi and Pasiouras (2010)), it is necessary to control for environmental characteristics to obtain 'fair' efficiency evaluation. There are multiple approaches for this sake[11]. The far-most popular approach is also the most controversial one: a two-stage approach. The two-stage approach estimates in a first phase non-parametrically the efficiency scores (most commonly by FDH or DEA). In a second phase, it explains the obtained estimates by a parametric regression. For bank branches, a two-stage approach is used by e.g. Paradi et al. (2011). Simar and Wilson (2007), Simar and Wilson (2011), "*2-stage DEA: Caveat Emptor*" and Johnson and Kuosmanen (2012) show rigorously that the second-stage inference is invalid in the thousands of studies that use a two-stage approach. A two-stage approach imposes that the attainable set of input×output does not depend on the environment (the so called 'separability assumption'). In addition, there are numerous estimation issues. As bank branch location is considered as an important quality attribute of banking technology (Das and Kumbhakar, 2012), it is hard to as-

[11] For a review see e.g. Daraio and Simar (2007, p. 96-100)

sume that the attainable sales, given inputs, do not depend on for example the number of persons passing by or the level of competition. Therefore, we do not impose separability between the input×output space and the environment. We estimate conditional efficiency as proposed by Daraio and Simar (2005).[12] By altering the conditional survivor function $S_{\mathbf{Y}|\mathbf{X}}$ to account for influences of (in our case discrete) environmental variables $\mathbf{Z} \in \Re^r$ as in (6.7), we can define $\lambda_m(\mathbf{x},\mathbf{y}|\mathbf{z})$ as in (6.8).

$$S_{\mathbf{Y}|\mathbf{X},\mathbf{Z}} = \text{Prob}(\mathbf{Y} \geq \mathbf{y}|\mathbf{X} \leq \mathbf{x}, \mathbf{Z} = \mathbf{z}). \tag{6.7}$$

$$\lambda_m(\mathbf{x},\mathbf{y}|\mathbf{z}) = \int_0^\infty \left[1 - (1 - S_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}(u\mathbf{y}|\mathbf{x},\mathbf{z}))^m\right] du. \tag{6.8}$$

In real life applications, it is not always possible to set $\mathbf{Z} = \mathbf{z}$ due to data limitations, especially when $\mathbf{Z}$ is multivariate. Kernel smoothing is needed. Whereas in the unconditional order-m procedure all units have an equal probability of being selected for membership in the sub sample Ψ_m, in the conditional order-m procedure, the probability to be drawn is defined on the basis of a kernel weighting function. The basic idea behind kernel smoothing is that observations with a similar operating environment get a greater probability to be drawn for membership in Ψ_m. Differently put, $\text{Prob}(\mathbf{Y} \geq \mathbf{y}|\mathbf{X} \leq \mathbf{x}, \mathbf{Z} = \mathbf{z})$ is estimated by $\text{Prob}(\mathbf{Y} \geq \mathbf{y}|\mathbf{X} \leq \mathbf{x}, \mathbf{Z}$ close to $\mathbf{z})$. To weight the discrete variable $\mathbf{z}_j$ with c_j categories, we define an Aitchison and Aitken (1976) discrete kernel weight function with localization parameter $\mathbf{h}^d$ as in (6.9). In the extreme case of $h_j^d = 0$, no weight is given to observations with a different value of Z. Only observations with $\mathbf{Z}_i j = \mathbf{z}_j$ are drawn for membership in Ψ_m. The other extreme of $h_j^d = (c_j - 1)/c_j$ means that observations with $Z_{ij} \neq z_j$ get equal weight to be drawn as observations with $Z_{ij} = z_j$. In other words, $\mathbf{Z}$ is ignored by the model. Values between the lower and upper bound indicate that more weight is given to observations with the same value of $\mathbf{Z}$ as the observation than to observation with a different value of $\mathbf{Z}$.

As is common practice, we make use of a product kernel to allow for multivariate kernel weighting: $L_{\mathbf{h}^d}(\mathbf{Z}_i, \mathbf{z}) = \prod_{j=1}^r l^d(Z_{ij}, z_j, h_j^d)$. A data-driven least-squares cross-validation procedure as

[12] Jeong et al. (2010) show the asymptotic properties of the conditional frontier approaches.

described in Badin et al. (2010) is used to select the optimal value of $\mathbf{h}^d$ in $\hat{S}_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}$ as defined in (6.10). Replacing $S_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}$ by $\hat{S}_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}$ in (6.8) gives us $\hat{\lambda}_m(\mathbf{x},\mathbf{y}|\mathbf{z})$.

$$l^d(Z_{ik}^d, z_k^d, h_k^d) = \begin{cases} 1 - h_k^d & \text{if } Z_{ik}^d = z_k^d, \\ h_k^d/(c_k - 1) & \text{otherwise.} \end{cases} \tag{6.9}$$

$$\hat{S}_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}(\mathbf{y}|\mathbf{x},\mathbf{z}) = \frac{\sum_{i=1}^n I(\mathbf{X} \leq \mathbf{x}, \mathbf{Y} \geq \mathbf{y}) L_{\mathbf{h}^d}(\mathbf{Z},\mathbf{z})}{\sum_{i=1}^n I(\mathbf{X} \leq \mathbf{x}) L_{\mathbf{h}^d}(\mathbf{Z},\mathbf{z})}. \tag{6.10}$$

To visualize the estimated influence of $\mathbf{Z}$ on the production process, we regress $\hat{\mathbf{Q}} = \hat{\lambda}_m(\mathbf{X},\mathbf{Y}|\mathbf{Z})/\hat{\lambda}_m(\mathbf{X},\mathbf{Y})$ on $\mathbf{Z}$ as proposed by Daraio and Simar (2005, 2007a):

$$\hat{Q}_i = f(\mathbf{Z}_i) + \varepsilon_i \text{ ,with } i = 1,...,n. \tag{6.11}$$

To avoid the imposition of an arbitrary assumption on the functional form, we estimate a non-parametric 'local-linear' regression. A local linear relation is estimated for each observation point by obtaining a and b in (6.12).

$$\min_{\{a,b\}} \sum_{i=1}^n (\hat{Q}_i - a - (\mathbf{Z}_i - \mathbf{z})'b)^2 L_{\mathbf{h}}(\mathbf{Z}_i,\mathbf{z}). \tag{6.12}$$

For the evaluation points $\{\mathbf{z}_1,...,\mathbf{z}_s,...,\mathbf{z}_t\}$, we estimate the fitted values $\hat{\pi}_m^{\mathbf{z}_s} = E[\hat{\mathbf{Q}}_m|\mathbf{Z} = \mathbf{z}_s]$. A $\hat{\pi}_m^{\mathbf{z}_s}$ that is higher (lower), holding everything else equal, indicates that the operation environment is more (less) favourable.[13]

6.3.2 Multivariate stochastic nonparametric conditional frontier

This section briefly reviews the estimation of a multivariate local maximum likelihood stochastic conditional frontier. Our overview starts from the Kumbhakar et al. (2007) model with univariate output and multivariate input. We discuss our nonparametric frontier approach in the presence of both multivariate inputs as well as multivariate outputs, following Simar and Wilson (2011), and

[13] We make use of the R package 'np' of Hayfield and Racine (2008) to estimate the CDF and local-linear regression.

proceed by describing how we condition this nonparametric stochastic frontier on environmental variables, as in Park et al. (2010). We end this section by describing how we estimate stochastic versions of conditional efficiency for individual observations. Full details can be found in Kumbhakar et al. (2007), Simar and Wilson (2011) and Park et al. (2010).

We consider a set of i.i.d. random variables $(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i)$, for $i = 1, ..., n$, with input $\mathbf{X}_i \in \Re_+^p$, output $Y_i \in \Re_+^1$ and discrete environmental variables $\mathbf{Z}_i \in \Re^r$. The local maximum likelihood is based on a local parametric anchorage model. Typically, the frontier function $r(\mathbf{X}, \mathbf{Z})$ is introduced as in the parametric model of Aigner et al. (1977):

$$\log Y_i = r(\mathbf{X}_i, \mathbf{Z}_i) - u_i + v_i \text{ , with } i = 1, ..., n. \tag{6.13}$$

The inefficiency term $\mathbf{u}$ is in this work specified to have a half normal distribution $(\mathbf{u} \sim |N(0, \sigma_{\mathbf{u}}^2(\mathbf{x}, \mathbf{z})|))$, the error term $\mathbf{v}$ is normally distributed $(\mathbf{v} \sim N(0, \sigma_{\mathbf{v}}^2(\mathbf{x}, \mathbf{z})))$ and $\mathbf{u}$ and $\mathbf{v}$ are independent conditionally on $(\mathbf{X}, \mathbf{Z})$.[14] The conditional pdf for $\mathbf{Y}$ given $(\mathbf{X}, \mathbf{Z})$: $pdf(\mathbf{y}|\mathbf{x}, \mathbf{z}) = g(\mathbf{y}, r(\mathbf{x}, \mathbf{z}), \tau(\mathbf{x}, \mathbf{z}))$, where $r(\mathbf{x}, \mathbf{z})$ and $\tau(\mathbf{x}, \mathbf{z})$ - which is the pair $(\sigma_{\mathbf{u}}^2(\mathbf{x}, \mathbf{z}), \sigma_{\mathbf{v}}^2(\mathbf{x}, \mathbf{z}))$ - are to be estimated and g is assumed to be known.

Simar and Wilson (2011) extended the stochastic frontier model which serves as anchorage model to allow for multivariate output. Multivariate output $\mathbf{Y}_i \in \Re_+^q$ can be defined in polar coordinates (ω, η) with modulus $\omega(\mathbf{y}) = \sqrt{\mathbf{y}^T \mathbf{y}} \in \Re_+$ and amplitude (angle) $\eta = arctan(\frac{\mathbf{y}^{j+1}}{\mathbf{y}^1}) \in [0, \pi/2]^{q-1}$ where $j = 1, ..., q-1$. The authors show that the stochastic frontier model which serves as anchorage model can be written in its multivariate analog with a univariate modulus ω as dependent variable (see (6.14)). In the following, we define $\mathbf{X}$ as the pair $(\eta, \mathbf{X})$.

$$\log \omega_i = r(\mathbf{X}_i, \mathbf{Z}_i) - u_i + v_i. \tag{6.14}$$

As in the univariate output case, the inefficiency term $\mathbf{u}$ is specified to be half normally distributed, the noise term $\mathbf{v}$ is normally distributed and $\mathbf{u}$ and $\mathbf{v}$ are independent conditionally on

[14] Note however that this assumption is hard. Smith (2008) shows that the stochastic frontier estimates are significantly biased if the error component dependence is incorrectly ignored.

$(\mathbf{X},\mathbf{Z})$. It is important to note that both inefficiency and noise are defined in the appropriate radial direction. In specific, the inefficiency term $1/exp(-u_i)$ indicates the radial increase in all outputs needed to reach the frontier $exp(r(\mathbf{X}_i,\mathbf{Z}_i))$.

The basic idea of the nonparametric stochastic frontier approach is to use a local polynomial approximation to estimate the unknown local factors $r(\mathbf{x},\mathbf{z})$ and $\tau(\mathbf{x},\mathbf{z})$, which determine the conditional log-likelihood function as defined in (6.15).

$$L(r,\tau;\mathbf{X},\mathbf{Z}) = \sum_{i=1}^{n} \log g(\omega_i, r(\mathbf{X}_i,\mathbf{Z}_i), \tau(\mathbf{X}_i,\mathbf{Z}_i)). \tag{6.15}$$

The choice of the order of polynomials is discussed in Park et al. (2008). The authors found by simulation that in local likelihood estimation, first order polynomials are preferred when the interest is in the fitted value and second order polynomial terms are preferred if the interest is in the first derivatives. As we are interested in returns to scale and inefficiency, but not in the effect of inputs on the inefficiency and noise distribution per se, we could use local quadratic polynomials for the frontier and a local linear approximation for the variance functions. However, preliminary analysis showed that the combination of high dimensionality in our model with quadric approximation for the frontier resulted in inaccurate estimates in our sample. To avoid this identification problem, which is analog to the multicollinearity problem in parametric regressions, we restrict the model to a localized Cobb-Douglass model (i.e., Local Linear Maximum Likelihood Estimation).[15]

We follow the approach of Park et al. (2010) in localizing the stochastic frontier model in direction of both $\mathbf{X}$ and $\mathbf{Z}$. As discussed above, localization implies that we do not impose that a parametric form of the frontier and homogeneity in $\varepsilon = \mathbf{v} - \mathbf{u}$ holds for all units, but only locally, for units with a similar operating environment.

Gaussian kernel weight functions l^c as defined in (6.16) for $\mathbf{x}_k$ are used to give more weight to observations near the observation point. Window widths $\mathbf{h}^c$ impose the window of localization. If the window is very large, all observations are considered to be similar and we return

[15]Local Linear Maximum Likelihood Estimation took over 2 days in R on a workstation (INTEL XEON Duo CPU X5690 with 3,47 GHz, 12 cores and 96 GB RAM)

to the parametric case with a linear frontier and no heterogeneity in ε. If the window width is small, only some observations are considered to be similar to the observation. Non-linearities in the frontier and heterogeneity in ε are allowed. To allow for multivariate $\mathbf{X}$, we define -as is common practice - a product kernel $K_{\mathbf{h}^c}(\mathbf{X}_i,\mathbf{x}) = \prod_{k=1}^{q}(h_k^c)^{-1} l^c((X_{ik}-x_k)/h_k^c)$.

$$l^c\left(\frac{X_{ik}^c - x_k^c}{h_k^c}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X_{ik}^c - x_k^c}{h_k^c}\right)^2}. \tag{6.16}$$

An Aitchison and Aitken (1976) kernel l_j^d is specified to weight discrete variable $\mathbf{z}_j^d$ (see (6.9)). The discrete variables are only used to weight the likelihood function and are not included in the anchorage model as it makes no sense to approximate discrete variables by polynomials. We use a discrete product kernel $L_{\mathbf{h}^d}(\mathbf{Z}_i,\mathbf{z}) = \prod_{j=1}^{r} l^d(Z_{ij},z_j,\mathbf{h}_j^d)$ to localize in the direction of multiple environmental variables. The localization of the conditional log-likelihood of the stochastic frontier model in direction of both $\mathbf{X}$ and $\mathbf{Z}$ is defined in (6.17). By maximization of the localized conditional log-likelihood function, $\hat{r}(\mathbf{x},\mathbf{z}) = \hat{r}_0 + \hat{r}_1(\mathbf{X}_i - \mathbf{x})$ and $\hat{\tau}(\mathbf{x},\mathbf{z}) = \hat{\tau}_0 + \hat{\tau}_1(\mathbf{X}_i - \mathbf{x})$ are obtained (see (6.18)).

$$\begin{aligned} &L_n(r_0, r_1, \tau_0, \tau_1; \mathbf{X}, \mathbf{Z}) \\ &= \sum_{i=1}^{n} \log\, g(\omega_i, r_0 + r_1(\mathbf{X}_i - \mathbf{x}), \tau_0 + \tau_1(\mathbf{X}_i - \mathbf{x})) K_{\mathbf{h}^c}(\mathbf{X}_i - \mathbf{x}) L_{\mathbf{h}^d}(\mathbf{Z}_i, \mathbf{z}). \end{aligned} \tag{6.17}$$

$$(\hat{r}_0, \hat{r}_1, \hat{\tau}_0, \hat{\tau}_1) = \arg \max_{r_0, r_1, \tau_0, \tau_1} L_n(r_0, r_1, \tau_0, \tau_1; \mathbf{X}, \mathbf{Z}). \tag{6.18}$$

The choice of multivariate bandwidth $\mathbf{h} = [\mathbf{h}^c, \mathbf{h}^d]$ is of crucial importance. We opt for the often used data-driven approach that minimizes the asymptotic integrated mean squared error (AIMSE): the least-squares cross-validation approach as defined in (6.19). We estimate an optimal value of h by least squares cross-validation for a wide grid of values of $\mathbf{h}$.[16]

[16] In a real life application, it is necessary to obtain very high precision of the results. Therefore, we did not use a subset of observations to perform the cross-validation procedure as in the methodological studies of Kumbhakar et al. (2007), Simar and Wilson (2011) and Park et al. (2010)

$$CV(\mathbf{h}) = \frac{1}{n}\sum_{i=1}^{n}((\log \omega_i - (\hat{r}_0^{(i)}(\mathbf{X}_i, \mathbf{Z}_i) - \hat{u}_i^{(i)}))^2 t, \tag{6.19}$$

where $\hat{r}_0^{(i)}$ and $\hat{u}_i^{(i)}$ are the leave-one-out version of the local linear estimators and $0 \leq t \leq 1$ is a trimming weight to limit the the sensitivity of the routine to potential numerical problems and outlying values.

Simulations of Simar and Wilson (2011) show that the distributional assumptions in the anchorage model are not restrictive for the model as a whole. However, the individual efficiency scores are highly influenced by the distributional assumptions on the convolution term $\varepsilon = \mathbf{v} - \mathbf{u}$. In this respect, Simar and Wilson (2011) propose to not decompose the convolution term $\varepsilon = \mathbf{u} - \mathbf{v}$. The authors propose a two-step procedure where in a first step the nonparametric stochastic frontier model pre-whitens the frontier and where in a second step stochastic versions of FDH/DEA estimators are derived. In other words, the authors propose to estimate $\widetilde{\lambda}_i = exp(\widetilde{u_i - v_i})$, where the wide tilde denotes that FDH (DEA) is used to obtain a (convex) free disposable hull of the estimated nonparametric frontier. In contrast to Simar and Wilson (2011), we condition on environmental variables. Therefore, we cannot use the unconditional efficiency measures FDH/DEA in a second step to monotonize and convexify the frontier. We define inefficiency directly as the proportional augmentation needed in all outputs to reach the pre-whitened conditional frontier $exp(r(\mathbf{X}_i, \mathbf{Z}_i))$, that is: $\hat{\lambda}_{NSF}(\mathbf{X}_i, \mathbf{Y}_i|\mathbf{Z}_i) = exp(\hat{u}_i - \hat{v}_i)$. We thus allow for non-monotonicity caused by for example congestion. $\hat{\lambda}_{NSF}(\mathbf{X}_i, \mathbf{Y}_i|\mathbf{Z}_i)$ can be seen as a stochastic estimator of conditional efficiency.

6.4 Results

We present the distribution across branches of our efficiency results in Table 6.8. The DEA efficiency scores in the first row of the Table are clearly lower on average than the efficiency scores from other methods, which points towards a distorting effect of outliers in the sample. We also observe that the FDH based approaches have more than 25% fully efficient branches, which is to be expected from the method and which is desirable from the point of view of our application:

the main focus is to identify with certainty a sufficiently large proportion of underperformers. We also observe that the distributions and the standard deviations of the two methods we focus on (last two rows of Table 6.8) are comparable, although the underlying assumptions on the distribution of noise are very different. Still, the correlations in Table 6.9 reveal how individual efficiency scores from methodologies of the same family are strongly correlated among each other (see the first four rows of Table 6.9), but that correlations are less strong with methods from the other methodological family (see the last row of Table 6.9), suggesting that the combination of results from our two methods will be informative.

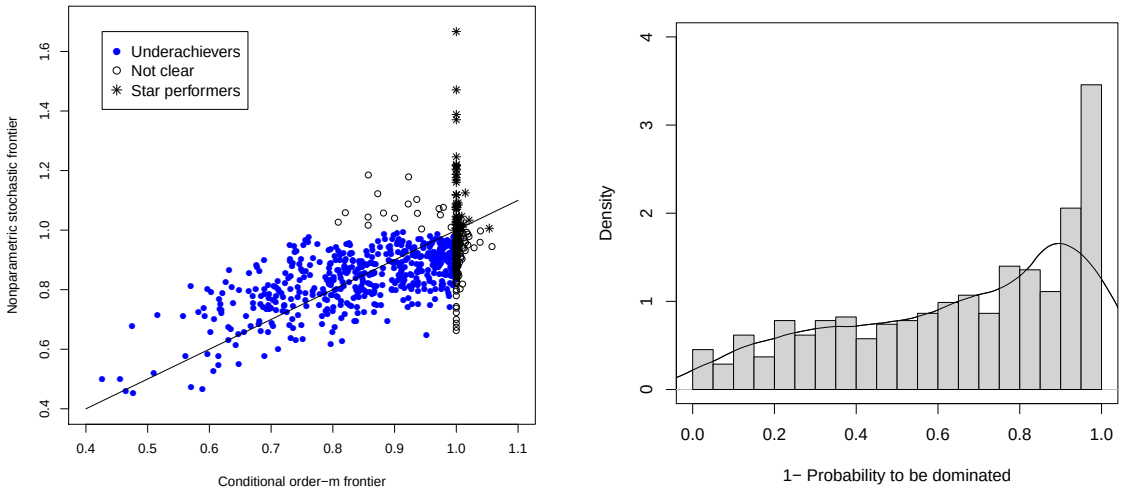
In Figure 6.3(a) we scatter plot the individual bank branch efficiency levels according to our two methods of choice. The method allows to very clearly identify a number of underachieving branches. So these bank branches are less efficient then their peers, controlling for their client profiles, region and urbanization typology and independent of the way how we disentangle noise from inefficiency. Our method even allows to rank the bank branches as in Figure 6.3(b). This Figure shows the distribution of the probability that a bank's efficiency is not lower than that of other underachievers $(1 - P(\hat{\lambda}_m(\mathbf{X},\mathbf{Y}|\mathbf{Z}) < \hat{\lambda}_m(\mathbf{x},\mathbf{y}|\mathbf{z})$ and $\hat{\lambda}_{NSF}(\mathbf{X},\mathbf{Y}|\mathbf{Z}) < \hat{\lambda}_{NSF}(\mathbf{x},\mathbf{y}|\mathbf{z}))$. The fat right tail of more than 30% of bank branches indicates that a considerable proportion are not underperformers. The left tail of the distribution are the true underachievers, that are very likely to do even worse than other underachievers. These are the branches that should get the first attention if the bank wants to rachet up its market efficiency. In this sense our method is very applicable to real life management challenges.

	Mean	St.Dev.	Min.	Q1	Med.	Q3	Max.
DEA $(1/\hat{\lambda}_{DEA}(\mathbf{x},\mathbf{y}))$	0.68	0.15	0.31	0.57	0.67	0.78	1.00
FDH $(1/\hat{\lambda}_{FDH}(\mathbf{x},\mathbf{y}))$	0.84	0.15	0.39	0.73	0.86	1.00	1.00
Order-m $(1/\hat{\lambda}_{m}(\mathbf{x},\mathbf{y}))$	0.88	0.16	0.41	0.77	0.91	1.00	1.44
Conditional order-m $(1/\hat{\lambda}_{m}(\mathbf{x},\mathbf{y}\|\mathbf{z}))$	0.89	0.12	0.43	0.81	0.93	1.00	1.06
Nonparametric SF $(1/\hat{\lambda}_{NSF}(\mathbf{x},\mathbf{y}\|\mathbf{z}))$	0.88	0.12	0.45	0.80	0.89	0.94	1.67

Table 6.8: Efficiency estimates

	DEA	FDH	Order-m	Cond. order-m	NSF
DEA	1.00	0.79	0.85	0.78	0.66
FDH	0.79	1.00	0.96	0.93	0.61
Order-m	0.85	0.96	1.00	0.92	0.64
Cond. order-m	0.78	0.93	0.92	1.00	0.63
NSF	0.66	0.61	0.64	0.63	1.00

Table 6.9: Correlogram of efficiency estimates



(a) Efficiency matrix

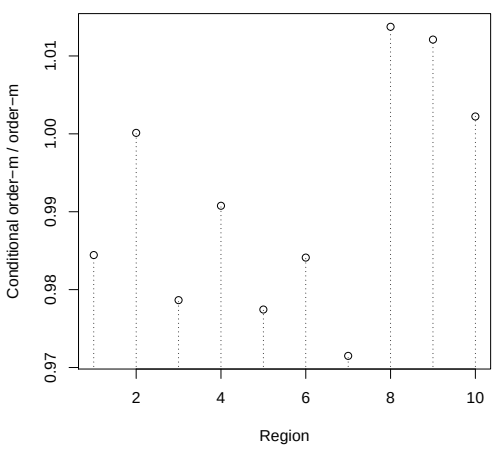
(b) Probability to not be dominated by other underachievers

Figure 6.2: Visualization of the ranking

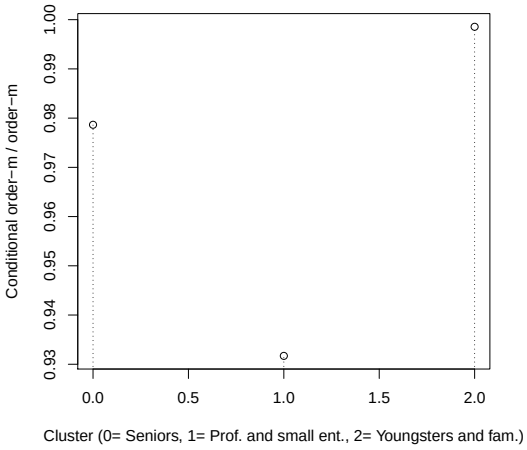
Figure 6.3 shows the estimated ratio between conditional order-m inefficiency $\hat{\lambda}_m(\mathbf{x},\mathbf{y}|\mathbf{z})/\hat{\lambda}_m(\mathbf{x},\mathbf{y})$ for the chosen evaluation points.[17] In all results, the value of the other discrete environmental variables is set to be equal to their respective modes. The modes are respectively 'Region 3', 'Seniors' and 'Well equipped non-urban city'. We observe immediately that the controls for

[17]We provide no confidence intervals as there is at this moment no published work on confidence intervals in a conditional efficiency framework.

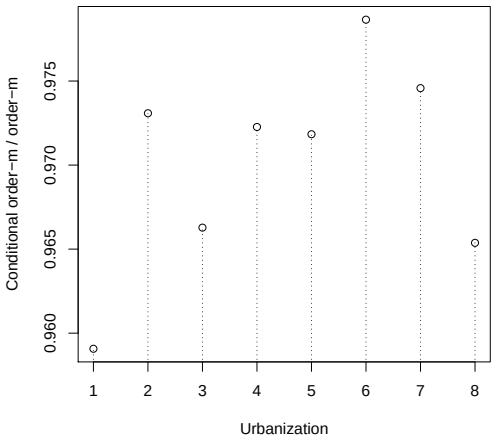
environmental factors have a noticeable effect on bank branch efficiency. Figure 6.3(a) shows that conditional order-m inefficiency of evaluated branches from region 7 is 3% lower than the unconditional order-m inefficiency, while it is clearly higher for regions 8 and 9. In other words, we find that it is indeed more difficult to operate in region 7 and more easy in regions 8 and 9. Effects of client profile are even more substantial. Figure 6.3(b) indicates that the conditional inefficiency of branches with a client profile of professionals and small enterprises is 7% lower than the unconditional inefficiency. Regarding the effects of urbanization typology, Figure 6.3(c) shows that it is more difficult to produce sales and loyalty in a large city, presumably because of more competititon and smaller switching costs. Failing to control for these environmental factors would therefore indeed have biased our efficiency estimates.



(a) Effect Region



(b) Effect Client profile



(c) Effect Urbanization

Figure 6.3: Visualization of the effects of environmental variables on the production environment

In Table 6.10 we show the basic elasticity estimates of the nonparametric stochastic frontier. We readily observe that on average returns to scale are positive. The nonparametric stochastic frontier results also give insight in the distribution of marginal scale elasticities in Figure 6.4. To evaluate the influence of a specific environmental variable on marginal returns to scale, we use the subsample of observations for which the other environmental variables are equal to their

modes. We find first of all considerable regional disparities in returns to scale (see Figure 6.5). In addition, almost all observations with a client profile of 'Youngsters and Families' and 'Seniors' are characterized by increasing returns to scale while branches with a client profile of 'Professionals and Small Enterprises' are characterized by decreasing returns to scale in the selected region and city type. We find no clear effect of urbanization typology in our subsample of units from region 3 and with a client profile of 'Seniors'.

	Mean	St.Dev.	Min.	Q1	Med.	Q3	Max.
Effect log staff in FTE	0.83	0.12	0.42	0.76	0.84	0.91	1.17
Effect log market potential	0.05	0.09	-0.17	-0.01	0.06	0.12	0.31
Effect log ATM's	0.17	0.10	-0.13	0.10	0.16	0.24	0.52
Returns to scale	1.06	0.11	0.65	0.99	1.05	1.13	1.32
Effect output mix	-0.30	0.26	-1.03	-0.47	-0.30	-0.12	0.48

Table 6.10: Elasticity estimates

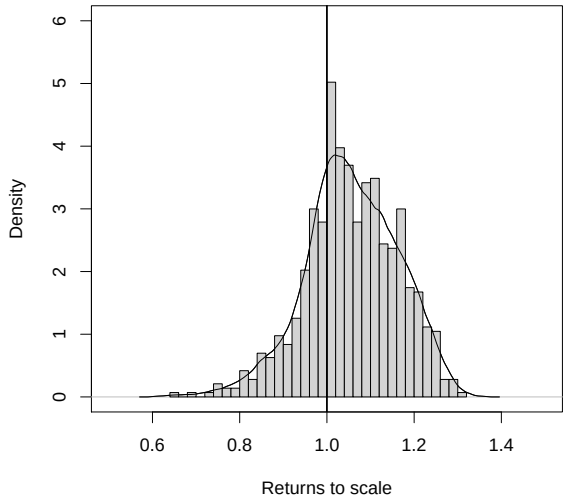
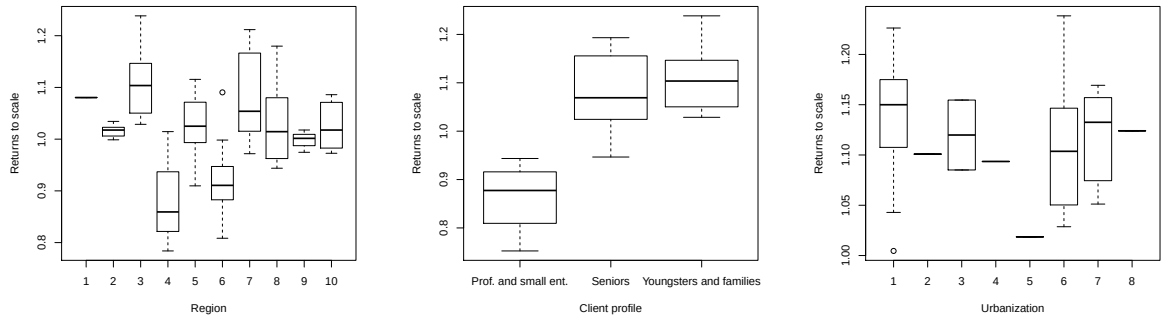


Figure 6.4: Returns to scale



(a) Effect Region (b) Effect Client profile (c) Effect Urbanization

Figure 6.5: Indications on how the environmental variables influence returns to scale

6.5 Conclusion

We estimate bank branch efficiency of a large bank in Belgium. Since it is not a priori clear how we should treat noise, we are the first to calculate efficiency and compare results according to two very different methodological approaches, namely the localized maximum likelihood stochastic frontier approach and the conditional order-m approach. Both methods have solved a lot of the problems of their predecessors, stochastic fronter analysis and Data Envelopment Analysis respectively, but still retain a very different approach to noise. We start from a multi-input multi-output model, where detailed information on market potential is used as an input and information on client loyalty is one of the outputs. We control all our estimations for important environmental variables, such as region specific effects, client profile and urbanization typology.

The result is 1) an unambiguous and fair classification of a sufficiently large proportion of the branches into the bin of underachievers and 2) a ranking of these branches. We indeed find that market potential plays a role, that environmental factors, like client profile, are very important, and that scale elasticities may vary over branches and depend on these environmental factors. The method also allows us to derive marginal effects of certain inputs, so that we can pinpoint the source of the low efficiency of a given branch. In this sense our method is very applicable to

real life challenges.

6.6 Appendix

6.6.1 Market potential

Figure 6.6 gives an overview of the construction of the confidential bank branch level index of market potential. As discussed above, a 'jar of 10,000,000 potential points' are distributed over all the communities/quarters in Belgium. To capture the potential for selling financial products to physical persons (retail market), the bank corrects for differences in average income, income classes (to correct for the fact that physical persons with higher income have proportionally more income available to buy financial products), the financial capacity of clients (liabilities and assets) and differences in the 'consumption-free gross revenue' between age cohorts. For this, city level NIS-data are combined with quarter-level bank data.

To capture the potential of professionals, the bank attributed 250,000 points to professionals and 1,000,000 points to self-employed. First, the bank corrects for the total income received per county and NACE-code (social security data). Second, the bank controls for differences between home and work address by 1) including directly the number of employees in a NACE-code in an county and 2) using data of the 'Social and Economic Survey 2001' that captures differences in home and work address for self-employed. To capture the potential of small enterprises, the bank uses social security data on enterprises with a maximum of 20 employees on value added and turnover per NACE code. To correct for differences in the importance of sectors, the sectors are weighted in function of the number of employees per sector in the specific city.

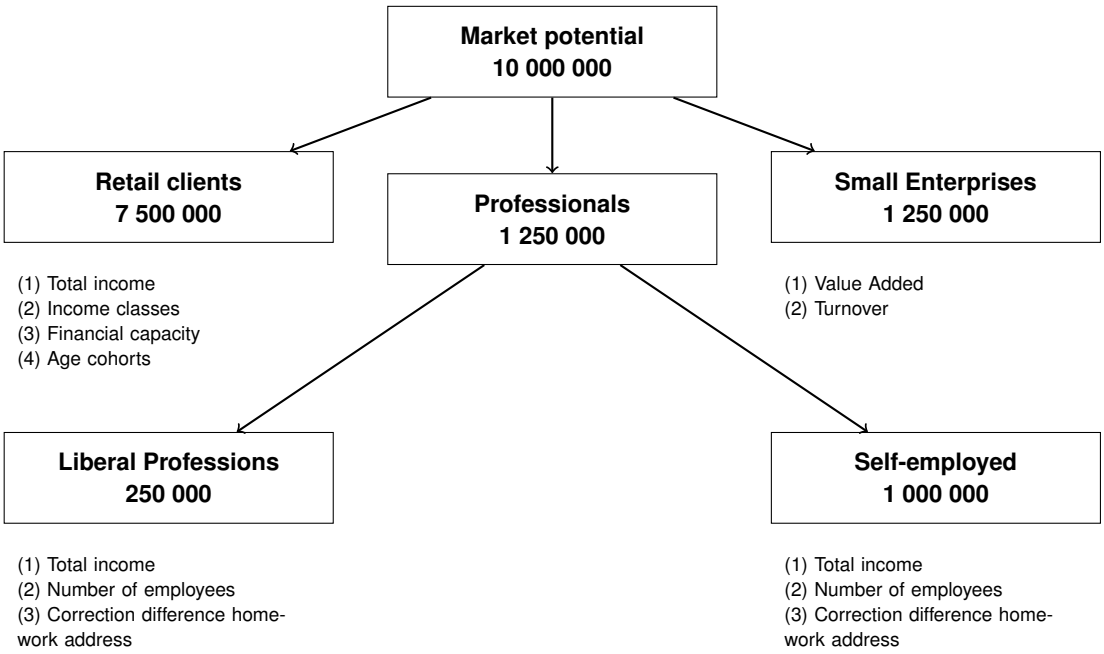


Figure 6.6: Market potential

6.6.2 Bandwidth sizes

Variable	Bandwidth
Staff in FTE	0.559
Market Potential	1.560
Number of ATM's	4.165
Output mix	1.368
Region	0.064
Client Profile	0.012
Urbanization	0.775

Table 6.11: Estimated optimal bandwidth sizes NSF model

	Mean	St.Dev.	Min.	Q1	Med.	Q2	Max.
Region	0.560	0.170	0.000	0.519	0.577	0.644	0.900
Client Profile	0.451	0.175	0.000	0.346	0.476	0.602	0.667
Urbanization	0.773	0.146	0.000	0.736	0.833	0.871	0.875

Table 6.12: Data-driven bandwidth sizes to compute conditional efficiency

Variable	Bandwidth
Region	0.259
Client Profile	0.063
Urbanization	0.617

Table 6.13: Bandwidth sizes to perform local-linear regression on Q_z

References

Aigner, D., Lovell, C., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. Journal of Econometrics 6, 21–37.

Aitchison, J., Aitken, C. G. G., 1976. Multivariate binary discrimination by kernel method. Biometrika 63 (3), 413–420.

Aragon, Y., Daouia, A., Thomas-Agnan, C., 2005. Nonparametric frontier estimation: A conditional quantile-based approach. Econometric Theory 21 (2), 358–389.

Athanassopoulos, A., 1998. Nonparametric frontier models for assessing market and cost efficiency of large-scale bank branch networks. Journal of Money, Credit, and Banking 30 (2), 173–192.

Badin, L., Daraio, C., Simar, L., 2010. Optimal bandwidth selection for conditional efficiency measures: A data-driven approach. European Journal of Operational Research 201, 633–640.

Berger, A., 2007. International comparisons of banking efficiency. Financial Markets, Institutions and Instruments 16, 119–144.

Berger, A., Leusner, J., Mingo, J., 1997. The efficiency of bank branches. Journal of Monetary Economics 40, 141–162.

Berger, A. N., Humphrey, D. B., 1997. Efficiency of financial institutions: International survey and directions for future research. European Journal of Operational Research 98 (2), 175–212.

Bos, J., Kool, C., 2006. Bank efficiency: the role of bank strategy and local market conditions. Journal of Banking & Finance 30, 1953–1974.

Bos, J. W. B., Koetter, M., Kolari, J. W., Kool, C. J. M., 2009. Effects of heterogeneity on bank efficiency scores. European Journal of Operational Research 195 (1), 251–261.

Camanho, A., Dyson, R., 2005. Cost efficiency, production and value-added models in the analysis of bank branch performance. Journal of the Operational Research Society 56, 483–494.

Cazals, C., Florens, J. P., Simar, L., 2002. Nonparametric frontier estimation: A robust approach. Journal of Econometrics 106 (1), 1–25.

Charnes, A., Cooper, W. W., Rhodes, E., 1978. Measuring efficiency of Decision-Making Units. European Journal of Operational Research 2 (6), 429–444.

Cook, W., Hababou, M., 2001. Sales performance measurement in bank branches. Omega 29, 299–307.

Cook, W., Hababou, M., Tuenter, H., 2000. Multicomponent efficiency measurement and shared inputs in Data Envelopment Analysis: An application to sales and service performance in bank branches. Journal of Productivity Analysis 14, 209–224.

Daouia, A., Gijbels, I., 2011. Robustness and inference in nonparametric partial frontier modeling. Journal of Econometrics 161 (2), 147–165.

Daouia, A., Ruiz-Gazen, A., 2006. Robust nonparametric frontier estimators: Qualitative robustness and influence function. Statistica Sinica 16 (4), 1233–1253.

Daouia, A., Simar, L., 2007. Nonparametric efficiency analysis: A multivariate conditional quantile approach. Journal of Econometrics 140 (2), 375–400.

Daraio, C., Simar, L., 2005. Introducing environmental variables in nonparametric frontier models: A probabilistic approach. Journal of Productivity Analysis 24 (1), 93–121.

Daraio, C., Simar, L., 2007. Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications. Studies in productivity and efficiency. Springer Science and Business Media.

Das, A., Kumbhakar, S., 2012. Productivity and efficiency dynamics in Indian banking: An input distance function approach incorporating quality of inputs and outputs. Journal of Applied Econometrics 27 (2), 205–234.

Debreu, G., 1951. The coefficient of resource utilization. Econometrica 19, 273–292.

Degryse, H., Ongena, S., 2001. Bank relationships and firm profitability. Financial Management 30, 9–34.

Degryse, H., Ongena, S., 2005. Distance, lending relationships, and competition. Journal of Finance 60 (1), 231–266.

Deprins, D., Simar, L., Tulkens, H., 1984. Measuring labor-efficiency in post offices. In: Marchand, M., Pestieau, P., Tulkens, H. (Eds.), The performance of public enterprises - concepts and measurement. Amsterdam, North-Holland, pp. 243–267.

Elyasiani, E., Goldberg, L., 2004. Relationship lending: A survey of the literature. Journal of Economics and Business 56 (4), 315–330.

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

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

Fethi, M., Pasiouras, F., 2010. Assessing bank efficiency and performance with operational research and artificial intelligence techniques:A survey. European Journal of Operational Research 204, 189–198.

Freixas, X., 2005. Deconstructing relationship banking. Investigaciones Economicas 29 (1), 3–31.

Hayfield, T., Racine, J. S., 2008. Nonparametric econometrics: The np package. Journal of Statistical Software 27 (5), 1–32.

Hirtle, B., 2007. The impact of network size on bank branch performance. Journal of Banking & Finance 31, 3782–3805.

Jeong, S., Park, B., Simar, L., 2010. Nonparametric conditional efficiency measures: Asymptotic properties. Annals of Operations Research 173 (1), 105–122.

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

Kumbhakar, S., Tsionas, E., 2008. Scale and efficiency measurement using a semiparametric stochastic frontier model: Evidence from U.S. commercial banks. Empirical Economics 34, 585–602.

Kumbhakar, S. C., Park, B. U., Simar, L., Tsionas, E. G., 2007. Nonparametric stochastic frontiers: A local maximum likelihood approach. Journal of Econometrics 137 (1), 1–27.

Meeusen, W., van Den Broeck, J., 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. International Economic Review, 435–444.

Ongena, S., Smith, D. C., 2001. The duration of bank relationships. Journal of Financial Economics 61 (3), 449–475.

Paradi, J., Rouatt, S., Zhu, H., 2011. Two-stage evaluation of bank branch efficiency using Data Envelopment Analysis. Omega 39, 99–109.

Paradi, J., Vela, S., Yang, Z., 2004. Assessing bank and bank branch performance: Modeling considerations and approaches. In: Cooper, W., Seiford, L., Zhu, J. (Eds.), Handbook on Data Envelopment Analysis. Kluwer Academic Publishers, pp. 349–400.

Park, B., Simar, L., Zelenyuk, V., 2010. Local maximum likelihood techniques with categorical data. Centre for Efficiency and Productivity Analysis WP14.

Park, B. U., Simar, L., Zelenyuk, V., 2008. Local likelihood estimation of truncated regression and its partial derivatives: Theory and application. Journal of Econometrics 146 (1), 185–198.

Petersen, M. A., Rajan, R. G., 1995. The effect of credit market competition on lending relationships. Quarterly Journal of Economics 110 (2), 407–443.

Portela, M., Thanassoulis, E., 2007. Comparative efficiency analysis of Portuguese bank branches. European Journal of Operational Research 177, 1275–1288.

Rajan, R. G., 1992. Insiders and outsiders - The choice between informed and arms-length debt. Journal of Finance 47 (4), 1367–1400.

Rajan, R. G., Zingales, L., 2003. The great reversals: The politics of financial development in the twentieth century. Journal of Financial Economics 69 (1), 5–50.

Serra, T., Goodwin, B., 2009. The efficiency of Spanish arable crop organic farms, a local maximum likelihood approach. Journal of Producticity Analysis 31, 113–124.

Sharpe, S. A., Sep. 1990. Asymmetric information, bank lending, and implicit contracts - A stylized model of customer relationships. Journal of Finance 45 (4), 1069–1087.

Simar, L., 2007. How to improve the performances of DEA/FDH estimators in the presence of noise? Journal of Productivity Analysis 28 (3), 183–201.

Simar, L., Wilson, P. W., 2007. Estimation and inference in two-stage, semi-parametric models of production processes. Journal of Econometrics 136 (1), 31–64.

Simar, L., Wilson, P. W., 2011. Two-stage DEA: Caveat emptor. Journal of Productivity Analysis 36 (2), 205–218.

Smith, M., 2008. Stochastic frontier models with dependent error components. Econometrics Journal 11 (1), 172–192.

Tibshirani, R., Hastie, T., 1987. Local likelihood estimation. Journal of the American Statistical Association 82, 559–567.

Yatchew, A., 1998. Nonparametric regression techniques in economics. Journal of Economic Literature 36, 669–721.