UNIVERSITEIT GENT

FACULTEIT ECONOMIE EN BEDRIJFSKUNDE

TWEEKERKENSTRAAT 2
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

Estimating and explaining efficiency in a multilevel setting: A robust two-stage approach

Kristof De Witte and Marijn Verschelde

July 2010

2010/657

D/2010/7012/28

Estimating and explaining efficiency in a multilevel setting: A robust two-stage approach*

Kristof De Witte† Marijn Verschelde‡

July 1, 2010

Abstract

Various applications require multilevel settings (e.g., for estimating fixed and random effects). However, due to the curse of dimensionality, the literature on non-parametric efficiency analysis did not yet explore the estimation of performance drivers in highly multilevel settings. As such, it lacks models which are particularly designed for multilevel estimations. This paper suggests a semi-parametric two-stage framework in which, in a first stage, non-parametric α efficiency estimators are determined. As such, we do not require any *a priori* information on the production possibility set. In a second stage, a semiparametric Generalized Additive Mixed Model (GAMM) examines the sign and significance of both discrete and continuous background characteristics. The proper working of the procedure is illustrated by simulated data. Finally, the model is applied on real life data. In particular, using the proposed robust two-stage approach, we examine a claim by the Dutch Ministry of Education in that three out of the twelve Dutch provinces would provide lower quality education. When properly controlled for abilities, background variables, peer group and ability track effects, we do not observe differences among the provinces in educational attainments.

Keywords: Productivity estimation; Multilevel setting; Generalized Additive Mixed Model; Education; Social segregation

JEL-classification: C14, C25, I21

*We sincerely thank Glenn Rayp for his helpful suggestions and insightful comments on an earlier draft of this paper.

†Top Institute for Evidence Based Education Research, Maastricht University, Tongersestraat 55, 6800 MD Maastricht, the Netherlands and Faculty of Economics and Business, University of Leuven, Naamsestraat 69, 3000 Leuven, Belgium. email: kristof.dewitte@econ.kuleuven.be.

‡I acknowledge financial support from the Fund for Scientific Research Flanders (FWO Vlaanderen). Corresponding author: SHERPPA, Department of General Economics, Ghent University, Tweekerkenstraat 2, 9000 Gent, Belgium. email: marijn.verschelde@ugent.be.

1 Introduction

Insights on the drivers of efficiency is crucial to improve the performance of observations. Indeed, the 'raw' efficiency estimates provide mostly case-specific information, while the prediction and explanation of the 'raw' efficiency scores allows for a generalization of the results. The semi-parametric Stochastic Frontier Analysis framework (SFA; Meeusen and Van Den Broeck, 1977) can easily deal with determining the sign and significance level of efficiency drivers, even in a multilevel framework. The latter is convenient as various applications are characterized by complex hierarchical data structures. For example in education, 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, 1986). However, in most cases the researcher does not have any *a priori* information on the underlying production technology (Yatchew, 1998). As such, the model is often wrongly specified, which results in biased estimations (Hjalmarsson et al., 1996).
On the contrary, nonparametric frontier techniques as Free Disposal Hull (FDH; Deprins et al., 1984) and Data Envelopment Analysis (DEA; Charnes et al., 1978) do not assume any *a priori* specification on the production function. However, the traditional DEA models lack statistical inference (e.g., computing standard errors) and procedures to smoothly include the exogenous environment. In addition, the DEA literature is silent about how to estimate and explain efficiency in a clustered, hierarchical design. Recently, statistical inference has been introduced (Simar and Zelenyuk, 2007), and attempts have been undertaken to incorporate and explain the exogenous environment. We briefly discuss two procedures.
A first popular procedure to examine the drivers of performance consists of a two-stage model in which the DEA efficiency estimates of the first stage are regressed on potential performance drivers. However, Simar and Wilson (2007) rigorously argued some issues in applying the two-stage model with DEA and FDH estimators in the first stage. These issues (discussed in Section 3) make the traditional two-stage models intricate. Nevertheless, Banker and Natarajan (2008) and McDonald (2009) discuss that two-stage approaches may be valid if the inputs are not (too much) correlated with the environmental variables. To do so, McDonald (2009) suggests the use of a quasi-maximum likelihood approach à la Papke and Wooldridge (1996).
A second popular procedure is the conditional efficiency approach, as introduced by Cazals et al. (2002) and Daraio and Simar (2005). It can smoothly include the exogenous characteristics without imposing a separability condition (i.e., the exogenous characteristics do not influence the inputs and outputs). Thanks to bootstrap based routines, the conditional efficiency model can additionally determine the significance level of the exogenous variables

(De Witte and Kortelainen, 2008). However, the procedure is not designed for a highly multilevel framework as it is not possible to include random effects and the inclusion of fixed effects implies that the multivariate bandwidth selection becomes dramatically time consuming with many discrete variables (Badin et al., 2010; De Witte and Kortelainen, 2008). In addition, including fixed effects implies the inclusion of dummy variables for the group variables. In result, no other group variable can be included in the analysis (otherwise perfect multicollinearity).

The lack of an appropriate technique which allows the empirical researcher to examine nonparametrically the drivers of performance in a highly multilevel setting, is a major constraint. The assumption that the set of observations are realizations of identically, independently distributed random variables is hard for the large majority of applications. Disregarding the clustered, hierarchical design of the data generating process can lead to biased inference (Wood, 2006; Raudenbush and Bryk, 2002). Circumventing this issue by assuming parametric specifications on the production frontier, or by reducing the number of levels is often not an option as, respectively, the researcher has totally no information on the specification of the frontier and a reduction of the levels intricates the analysis. For example, in the performance assessment of students one wants to allow for random class and/or school effects. Reducing the number of schools or specifying an assumed production function for student attainments does not make sense. We could make a similar observation for other settings (e.g., in a performance comparison of regional entities (e.g., countries or municipalities) one wants to account for random regional disparities).

As such, this paper contributes to the literature in two perspectives. Firstly, by combining established frameworks from the nonparametric and semi-parametric literature and by carefully exploiting insights of these models, we propose 'a Robust two-stage model'. The traditional two-stage models estimate a deterministic frontier model in the first phase, and a semi-parametric bootstrap in a second phase. The suggested robust two-stage model uses robust α-efficiency estimates (Daraio and Simar (2007a) and Aragon et al. (2006); see Section 2) in the first phase, and regresses the outcomes on the set of exogenous variables by a Generalized Additive Mixed Model (GAMM) in the second stage. This raises three thoughts. Firstly, by using the robust α estimator in the first phase we carefully avoid the Simar and Wilson (2007) critique, which focused on the use of the deterministic DEA estimators in the first phase (see section 2 and 3 for a detailed discussion). Secondly, because of the semiparametric nature of the GAMM procedure, no parametric assumptions on the functional form of the second stage regression are imposed. Only a mild (and testable) additivity assumption

is needed. Thirdly, the GAMM procedure is required in the second phase as it is extremely flexible. Moreover, it can easily include random group effects, which in turn can be introduced to estimate a multilevel (= mixed) regression. As an additional advantage, the GAMM model allows for *Quasi* approaches which can be used to allow for over- or underdispersion in the error structure (i.e., the variability of the data is higher (lower) than expected from the statistical model). The validity of the proposed robust two-stage model is illustrated on a simulated data set in Section 4.

The second contribution arises from the application. In particular, we examine the claim of the Dutch Ministry of Education and Culture in that the school quality is lower in the Dutch provinces of Groningen, Friesland and Drenthe in comparison to the remaining 9 provinces. Whereas the Ministry considered school quality mainly as providing 'care' for the student, we estimate school quality as the ability to obtain educational attainments. We apply the robust two-stage model on a large sample of Dutch secondary school students. In doing so, we control for student abilities, student background characteristics, peer effects, and random school and class effects. Our results indicate the existence of strong peer effects, large effects of social segregation and the lack of provincial differences among student attainments.

The remainder of this paper is structured as follows. Section 2 discusses some concepts in estimating efficiency. Insights in these concepts are necessary to understand the issues of the traditional two-stage model. Therefore, Section 2 makes a clear distinction between the traditional frontier models and the novel robust estimation techniques. Section 3 explores with the traditional two-stage model and the conditional efficiency approach two popular techniques to examine the drivers of efficiency. As both techniques are not constructed for highly multilevel settings an alternative technique is proposed in Section 4, which develops the robust two-stage model. While Section 5 illustrates the appropriate working of the proposed framework, Section 6 examines the claim of the Dutch Ministry in that some provinces are providing lower educational quality. Finally, Section 7 concludes the paper.

2 A first stage: Estimating efficiency

2.1 The nonparametric (deterministic) frontier approach

This section explores some concepts and recent models in the nonparametric efficiency literature. Although the outline is limited to the output-oriented case, the extension to input-orientation is straightforward. Assume that producers use a heterogeneous non-negative input vector $x \in \Re_+^p$ to produce a heterogeneous output vector $y \in \Re_+^q$. The production set Ψ of

feasible input-output combinations can be defined as:

$$\Psi = \left\{ (x, y) \in \Re_{+}^{p+q} \mid \text{x can produce y} \right\}. \tag{1}$$

In estimating Ψ, two different strands has been developed. We discuss briefly (1) the traditional *full* frontier estimators and (2) the *robust* frontier estimators.

Firstly, the traditional 'Data Envelopment Analysis' (DEA; Charnes et al., 1978) literature[1] estimates the production set while including all observed input-output combinations. As such, it estimates the efficiency of observations relatively to a full frontier. Farell (1957) and Debreu (1951) were the first to acknowledge that the output-efficiency score (i.e., maximization of outputs y given the observed inputs x) of an observation (x, y) can be obtained as:

$$\lambda(x, y) = \sup\{\lambda | (x, \lambda y) \in \psi\}. \tag{2}$$

A value $\lambda(x, y) = 1$ indicates full technical efficiency (i.e., there are no observations which are able to produce more outputs for the given input set). A $\lambda(x, y) > 1$ indicates inefficiency, i.e., it is possible to have a radial increase of $\lambda(x, y)$ 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:

$$y^{\,\partial}(x, y) = \lambda(x, y)y. \tag{3}$$

Under the assumption of free disposability[2], probability theory can be used to interpret the efficiency scores. In particular, efficiency can be viewed as the proportional augmentation of output that unit $(x, y) \in \psi$ needs to obtain in order to have a zero percent probability to be dominated, given the inputs x. Following Cazals *et al.* (2002), this can be algebraically expressed as:

$$\lambda(x, y) = sup\{\lambda | S_{Y|X}(\lambda y|x) > 0\}, \text{ with } S_{Y|X} = \text{Prob}(Y \geq y | X \leq x). \tag{4}$$

By replacing in (4) the survival function $S_{Y|X}$ by its empirical version $\hat{S}_{Y|X}$, Free Disposal Hull (FDH) inefficiency estimates $\hat{\lambda}_{FDH}(x, y)$, as introduced in Deprins et al. (1984), are obtained. If additionally to FDH a convexity assumption is imposed, one obtains the Data Envelopment Analysis (DEA) inefficiency estimates $\hat{\lambda}_{DEA}(x, y)$.

Secondly, a more novel procedure to estimate the production set was introduced by Cazals et al. (2002). As the traditional FDH and DEA estimators are sensitive to outlying observations, this 'partial frontier approach' does no longer compute the efficiency scores relative

[1] As is common practice, we refer to 'the DEA literature' to point to the broad set of frontier estimation techniques.

[2] i.e., if $(x, y) \in \psi$, then any (x', y') such that $x' \geq x$ and $y' \leq y$ is also in ψ.

to all observations in the production set. In particular, the robust order-α quantile frontier approach of Aragon et al. (2005) considers efficiency as the proportional augmentation of output that unit $(x, y) \in \psi$ needs to have to obtain a 1-α percent probability to be dominated, given the inputs. By definition, α is in $[0, 1]$ and should be close to 1.

$$\lambda(x, y) = sup\{\lambda | S_{Y|X}(\lambda y|x) > 1 - \alpha\}, \text{ with } S_{Y|X} = \text{Prob}(Y \geq y | X \leq x) \tag{5}$$

As proven by Cazals et al. (2002) and Daouia and Simar (2007), partial frontier estimates converge by a $\sqrt{n}$-rate to the true partial frontier. Because the distance to a partial frontier - close to the full frontier - is estimated, partial frontier efficiency estimates are not bounded at one (as the traditional DEA models).

2.2 The domination approach: α efficiency

Alternatively to the nonparametric (deterministic) frontier approach, which estimates efficiency as a distance to a frontier, Daraio and Simar (2007a) and Aragon et al. (2006) propose to estimate efficiency as the probability that another observation does not produce more output with less or equal inputs. The obtained estimator is labeled as 'α efficiency'.[3] Technically, α efficiency estimates the order-α of the estimated quantile frontier which passes through this unit:

Definition $\alpha^{output}(x, y) = 1 - S_{Y|X}$, with $S_{Y|X} = \text{Prob}(Y \geq y \,| X \leq x)$.

The empirical estimation of the α efficiency is then obtained by estimating:

$$\hat{\alpha} = 1 - \hat{S}_{Y|X,n} + \frac{1}{N_x} = 1 - \frac{\sum_{i|X_i \leq x} I(Y_i \geq y)}{N_x} + \frac{1}{N_x}, \text{ with } N_x = \sum_{i=1}^{n} I(X \leq X_i). \tag{6}$$

This discrete and easy-to-use estimator of α efficiency in equation (6) has been proposed by Daraio and Simar (2007a). Aragon et al. (2006) extended this to a smooth and monotone estimator. For the ease of explanation, in this paper, we focus on the discrete estimator. However, our results can straightforwardly be extended to the smooth estimator.
The α efficiency has some attractive characteristics which makes it an appealing efficiency estimation procedure. We briefly present the appealing characteristics in six properties. Firstly and similar to the traditional methods, it relies on the monotonicity property. Indeed, under Assumption 2.1, α is monotone nonincreasing with x and monotone nondecreasing with y.

Assumption 2.1 *Assume that $S_{Y|X}(y|x)$ is continuous for any x. Then for (x, y) such that $S_{Y|X}(y|x) < 1$, $S_{Y|X}(y|x)$ is monotone nondecreasing with x and monotone nonincreasing with y.*

[3] Which is not the same as the order-α estimator of Daouia and Simar (2007).

Proposition 2.2 *Under Assumption 2.1, whenever defined, α is monotone nonincreasing with x and monotone nondecreasing with y.*

By construction, $S_{Y|X}(y|x)$ is monotone nonincreasing in y. The assumption that $S_{Y|X}(y|x)$ is monotone nondecreasing in x is a reasonable assumption in production analysis (as discussed in Daouia and Simar, 2007). The latter assumption states that there is more probability to observe a level of output higher than a fixed value y for firms using less input than a level of x_2, than for firms using less input than a level of input x_1 with $x_1 \leq x_2$ (Daouia and Simar, 2007). If the assumption seems invalid in a particular setting, Aragon et al. (2006) proposed to use isotonization in order to enforce monotonicity with respect to x.

The use of $\hat{\alpha}$ as an estimator of α requires that $\hat{\alpha}$ is a consistent and robust estimator.

Proposition 2.3 *$\hat{\alpha}$ is a consistent estimator of α: When $n \to \infty$, $\hat{\alpha} \to \alpha$.*

By construction, $\hat{\alpha}$ is a consistent estimator of α.[4] Unlike efficiency estimates with DEA or FDH, α estimates are not downward biased by construction.

Proposition 2.4 *$\hat{\alpha}$ is a robust estimator of α: for a given number of outliers, when $n \to \infty$, $\hat{\alpha}_{with\ outliers} \to \alpha$.*

The α efficiency estimator shares the robust properties of robust nonparametric efficiency estimation methods. In contrast to DEA or FDH, the impact of a given number of outliers decreases dramatically if the sample size increases. Proposition 2.3 and 2.4 are illustrated numerically in the Appendix.

Proposition 2.5 *α is distributed over [0,1] and $\hat{\alpha}$ is distributed over the interval $]0, 1]$.*

Proposition 2.5 indicates that α and $\hat{\alpha}$ are conveniently distributed over a fixed interval. An efficiency score equal to 1 denotes full efficiency.

Proposition 2.6 *$\hat{\alpha} \to \alpha$ at $\sqrt{n}$ convergence speed.*

In contrast to the traditional methods of DEA and FDH, α efficiency does not require frontier smoothing (which is required in (robust) DEA and FDH to estimate the frontier). Consequently, $\sqrt{n}$ convergence speed is preserved, where n denotes the units with $X \leq X_i$.

[4]This is because $\hat{S}_{Y|X}(y|x)$ is a consistent estimator of $S_{Y|X}(y|x)$ and $1/Nx \to 0$ if $n \to \infty$.

3 A second stage: Explaining efficiency

Typically a researcher is less interested in the efficiency scores *an sich*, but the more in the determinants behind those efficiency estimates. Indeed, while the efficiency scores can provide insights in the relative performance of an individual observation, the prediction and explanation of the scores allow for a more generalized interpretation. The literature counts various approaches to examine the efficiency drivers (for a recent overview, see De Witte and Kortelainen, 2008). This subsection explores with the two-stage approach and the conditional efficiency estimates two popular procedures.

3.1 Two-stage approach

The two-stage approach estimates in a first phase nonparametrically the efficiency scores (most commonly by FDH or DEA). In a second phase, it explains the obtained estimates by a (semi)parametric approach. Explaining efficiency in a two-stage approach consists thus of estimating (7) by (8):

$$\lambda_i = m(Z_i) + \epsilon_i \tag{7}$$

$$\hat{\lambda_i} = \hat{m}(Z_i) + \hat{\epsilon}_i. \tag{8}$$

Although the two-stage approach has been applied frequently, Simar and Wilson (2007) indicate rigorously that it can be a rather tricky procedure. In particular, they 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.[5] Three problems are addressed in Simar and Wilson (2007). (1) Traditional non-parametric efficiency estimates $\hat{\lambda}_i^{\text{DEA-FDH}}$ are downwards biased. (2) Efficiency estimates are serially correlated, in a complicated, unknown way. (3) In any meaningful two-stage analysis, the explanatory variables Z_i are correlated with the input-output set (X_i,Y_i). Consequently, Z_i is correlated with the error term ϵ_i.
Out of the three issues, the first one (i.e., the bias of the estimated $\hat{\lambda}_i^{\text{DEA-FDH}}$) leads in Maximum Likelihood Estimations to biased and inefficient estimates. However, this small sample problem can be mitigated by constructing bias corrected estimates by the implementation of an appropriate bootstrap algorithm (Simar and Wilson, 2007).
The latter two issues disappear asymptotically. However, they disappear only at the slow convergence rate of the traditional nonparametric efficiency estimates $\hat{\lambda}_i^{DEA-FDH}$ and not at the usual $\sqrt{n}$-rate achieved by MLEs in standard parametric (truncated) regression models (Simar and Wilson, 2007). Conventional inference is based on a second-order Taylor approximation. Higher-order terms cannot be disregarded because of the slow rate with which

[5] It is important to stress that the Simar and Wilson (2007) critique particularly originates from the use of full frontier estimates, as FDH and DEA, in the first stage. The next section draws further on this idea.

the serial correlation among the efficiency estimates disappears. Consequently, conventional inference will be invalid in this setting (Simar and Wilson, 2008). 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.

On the other hand, Banker and Natarajan (2008) show that a typical DEA two-stage approach is statistically meaningful and can be considered as a DEA-based stochastic frontier estimation. Banker and Natarajan (2008) and McDonald (2009) show that a typical DEA two-stage approach with OLS in the second stage yields consistent estimates if the inputs are not (too much) correlated with the environmental variables. Because OLS is not an asymptotically efficient estimator and because the predicted efficiency levels may not lie in the interval of true efficiency, McDonald (2009) propose to use a quasi-maximum likelihood approach à la Papke and Wooldridge (1996) to consistently and robustly estimate efficiency in a typical DEA two-stage approach. A logit transformation can be used to transform the unbounded linear predictors in the fixes efficiency interval.

Nevertheless, four important issues might arise in a two-stage approach with either full frontier DEA estimates in the first stage, or with bootstrapped DEA estimates in the first stage. (1) By the use of full frontier approaches in the first stage, the two-stage estimates share the vulnerability for outliers in X and Y. (2) A two-stage approach implies the introduction of a separability assumption between the $input \times output$ space and the space of Z values (Daraio and Simar, 2007a). (3) The two discussed approaches impose parametric assumptions on the functional form of the regression and error distribution. (4) The bootstrap algorithms and typical two-stage approaches are defined for and tested in 1-level settings.

Out of the four issues, the literature has already solved two. The separability assumption and parametric assumptions on the regression (issue (2)) can be avoided by the use of a conditional efficiency approach, introduced by Daraio and Simar (2005) and Daraio and Simar (2007b). Parametric assumptions in the second stage (issue (3)) can be avoided by the use of a truncated local likelihood two-stage approach as in Park et al. (2008). To robustly and flexibly explain efficiency in a two-stage approach with a multilevel setting, we propose a new, flexible and more robust approach to overcome the first, third and latter discussed caveat. This robust approach is discussed in Section 4.

3.2 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 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* × *output* space and the space of Z values. In other words, Z can influence the attainable set Ψ or the position of the frontier of the attainable set. The conditional survival function is expressed as:

$$S(x, y|z) = Prob(Y \geq y|X \leq x, Z = z), \tag{9}$$

such that the conditional efficiency is obtained as:

$$\lambda(x, y|z) = sup(\lambda|S_{Y|X,Z}(\lambda y|x, z) > 0). \tag{10}$$

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

The conditional efficiency approach shows its usefulness to robustly explain efficiency in a growing number of research papers (for an overview of about 18 papers see De Witte and Kortelainen, 2008). However, it is not advisable to use a conditional efficiency approach in a highly multilevel setting. In a multilevel setting, the inclusion of multiple levels implies the introduction of fixed group effects or random group effects. Including fixed group effects in a conditional efficiency approach is possible, but would imply (1) a dramatically high loss of degrees of freedom and (2) loss of information of group variables (in education, examples are class size, school type, school autonomy). To our best knowledge, the conditional efficiency approach is not suited to include random group effects.[6] Consequently, to explain efficiency in a robust approach in a highly multivariate or multilevel setting, an alternative semiparametric approach with random group effects is needed.

4 Combining insights: the robust two-stage approach

As both the existing two-stage approaches and the conditional efficiency model are inappropriate to use in a (highly) multilevel setting, this section proposes an alternative approach.

[6] Although De Witte and Kortelainen (2008) proved that the conditional efficiency model could include many discrete exogenous variables without influencing the curse of dimensionality, in practice, estimations with many discrete variables (say, more than 20) becomes impractical as the estimation time to estimate the Kernel bandwidth increases dramatically.

The proposed model uses insights from the previous sections and smoothly tackles the raised issues.

4.1 Introducing robust efficiency estimates in a two-stage approach

The critique of Simar and Wilson (2007) on the (traditional) two-stage model arises from the use of frontier techniques as FDH and DEA in the first stage. Indeed, FDH and DEA lead to biased and serially correlated estimates which are correlated with Z. Nevertheless, the description in the previous section on the robust efficiency estimates indicated that the robust efficiency models do not suffer from these issues. Therefore, if partial frontier estimates are used in the first stage, no bootstrap algorithm is needed for valid inference (as is also discussed in De Witte and Kortelainen, 2008).

Proposition 4.1 *Inference is asymptotically valid in a two-stage approach, when robust non-parametric frontier approaches such as the order-m frontier approach of Cazals et al. (2002) or the order-α quantile frontier approach of Aragon et al. (2005) are used in the first-stage.*

As discussed in the previous section, it is shown by Cazals et al. (2002) and Daouia and Simar (2007) that partial frontier estimates converge at $\sqrt{n}$-rate to the true partial frontier, this with n the units with $X \leq X_i$. Intuitively, this attractive property is obtained as there is no smoothing of the frontier. In result, both the correlation among the ϵ_i, as well as the correlation between ϵ_i and Z_i disappear at $\sqrt{n}$-rate. Consequently, inference in the second stage is asymptotically valid if robust estimators are used in the first stage.[7]
However, introducing robust non-parametric frontier methods in a two-stage procedure is not without problems because the robust efficiency estimates are not bounded at 1. By consequence, a meaningful 1-to-1 transformation is needed to map the unbounded linear predictors in the unknown interval [0, close to but larger than 1]. As to our best knowledge a similar approach is not available in the literature, the next subsection develops a proposal.[8]

4.2 a Robust two-stage approach

To allow for a valid robust inference in the second stage of a two-stage approach and to relax the strong assumption that the sample observations (x_i, y_i, z_i) in the observation set $\varsigma_n = \{x_i, y_i, z_i\}_{i=1}^n$ are realizations of identically, independently distributed random variables,

[7]It should, however, be noted that asymptotic results are achieved when the sample size goes to infinity and the proportion of observations with few $X \leq X_i$ in the whole sample goes to zero. This is a stronger condition than in traditional asymptotic theory.

[8]For example, if standardization is used to transform the efficiency estimates, inference is possible, but an interpretation of the coefficients that holds over different samples is lost.

we propose to use the following two stages:

Stage 1 Estimate for each observation (x, y) the α efficiency score of Daraio and Simar (2007a) or Aragon et al. (2006).
Stage 2 Explain the α efficiency scores by using a semiparametric Generalized Additive Mixed Model (GAMM). (See appendix for an overview of the applied GAMM approach.)

In contrast to previous suggested modeling techniques (as Tobit or bootstrap), the GAMM model is extremely flexible, and, as such, better suited as a second stage model. In particular, (1) both discrete and continuous variables can easily be introduced. (2) *A priori*, no functional relationship between the continuous variables and $\hat{\alpha}$ are imposed (only a basic additivity assumption is required). (3) Random group effects can be introduced to estimate a multilevel (= mixed) regression. (4) *Quasi* approaches can be used to allow for over- or underdispersion in the error structure.
Overdispersion (underdispersion) occurs when the variability of the data is higher (lower) than we would expect from the given statistical model. In the binomial case, overdispersion occurs when $var(\hat{\alpha}) > \mu(1-\mu)$, with $\mu = E[\hat{\alpha}]$. Underdispersion when $var(\hat{\alpha}) < \mu(1-\mu)$. Over- or underdispersion can be the result of the unknown serial correlation between efficiency estimates - as discussed in Simar and Wilson (2007). By this dependency, the independence assumption of the binomial model is violated. In addition, underdispersion is expected because the estimates are conditional on the set with $X \leq X_i$ instead of the whole sample - as in binomial models. If inefficiency and noise are relatively low - such that the observations are close to the frontier - a high proportion of values close to 1 can be expected. A quasi-binomial approach introduces an unknown scale parameter s such that $var(\hat{\alpha}) = s\mu(1-\mu)$. By the use of a quasi-binomial specification, estimates are more robust for unobserved heterogeneity and dependency.

Where the mapping of the unbounded linear predictors in the unknown interval
[0, close to but larger than 1] was a major issue before (see previous section) - inspired by Venables and Dichmont (2004) - we suggest to explain the α efficiency estimates by generalized model with a logit link which maps the unbounded linear predictors in the closed interval $[0, 1]$. As discussed above, over- or underdispersion can be expected in the robust two-stage approach. Therefore, the more flexible quasi-binomial error structure is chosen to allow for over- or underdispersion in the mean-variance relationship.

Analogous to Venables and Dichmont (2004), the GAMM model is specified as:

$$\hat{\alpha}|\zeta \sim \text{Quasi-binomial}\left(\mu = \frac{e^{\eta}}{1+e^{\eta}}, Var(\mu) = \frac{\mu(1-\mu)}{T/s}\right),$$
$$\text{with } \eta = \beta_0 + \beta_1(z_1) + \beta_2(z_2) + \ldots + \beta_p(z_p) + s_1(z_{p+1}) + s_2(z_{p+2}) + \ldots + s_q(z_{p+q}) + W\zeta \tag{11}$$

where $\hat{\alpha}$ denotes the estimated α efficiency, $E[\hat{\alpha}] = \mu$, η the predictor, Z the fixed effects variables with p discrete variables and q continuous variables, W the random effects variables, β the coefficients for Z, ζ the multivariate random effects, T a weights parameter and s the scale parameter to allow for over- or underdispersion.

The estimation of this type of GAMM is implemented by the 'GAMM' function in the package 'mgcv' in R. A detailed description of the method can be found in Wood and Augustin (2002), Wood (2004) and Wood (2006). Particular features of the 'mgcv' package are that (1) a variety of approaches can be used for automatic smoothing, (2) interactions between variables that are smoothed (smooth terms) can be included, (3) mixed models are supported by the program[9], (4) confidence intervals of the smooth terms can be constructed by the use of Bayesian approximation, (5) the number of degrees of freedom for the smooth terms that are justified by the data can be estimated by the use Bayesian approximation, (5) estimation is computationally efficient. In what follows, penalized regression splines - as discussed in appendix and Wood (2006)- are used to estimate the smooths. This approach controls the model's smoothness by adding a wiggliness penalty λ to the least squares fitting problem. By the automatic estimation of λ - by minimizing a (generalized) cross-validation function - the selection of an appropriate degree of smoothing is automatic.[10]

5 Numerical illustrations

To illustrate the appropriate working of the proposed robust two-stage approach, we simulate data following Badin et al. (2010). We simulate in three steps: (1) a two-variate one-level setting, (2) a two-variate multilevel setting and (3) a multivariate multilevel setting. Our results clearly indicate the appropriate working of the suggested procedure.

[9]For this, it is assumed that the interest of the researcher is primarily on the fixed effects, including the smooth terms. Random effects and correlation structures are estimated to structure the residual correlation in function of the estimation of the fixed effects (Wood, 2006).

[10]The R code is available upon simple request from the authors.

5.1 Step 1: Explaining efficiency in a one-level setting

We simulate a convex technology with 2 additive outputs and 2 inputs. The efficient frontier is defined as:

$$y^{(2)} = 1.0845(x^{(1)})^{0.3}(x^{(2)})^{0.4} - y^{(1)}, \tag{12}$$

where $y^{(1)}$, $y^{(2)}$, $x^{(1)}$ and $x^{(2)}$ are components of, respectively, y and x. We generate $X_i^{(j)} \sim U(1,2)$, $\tilde{Y}_i^{(j)} \sim U(0.2,5)$. The output efficient random points on the frontier are simulated as:

$$Y_{i,eff}^{(1)} = \frac{1.0845(X_i^{(1)})^{0.3}(X^{(2)})^{0.4}}{S_i + 1} \tag{13}$$

$$Y_{i,eff}^{(2)} = 1.0845(X_i^{(1)})^{0.3}(X^{(2)})^{0.4} - Y_{i,eff}^{(1)} \tag{14}$$

where $S_i = \tilde{Y}^{(2)}/\tilde{Y}^{(1)}$ and denotes the slopes which characterize the generated random rays in the output space for j=1,2.

To introduce inefficiency and environmental variables in a 1-level setting, we generate $U_i \sim Expo(mean = 1/2)$ and $Z_j \sim U(1,4)$ for j=1,2. Output is defined by the following production process:

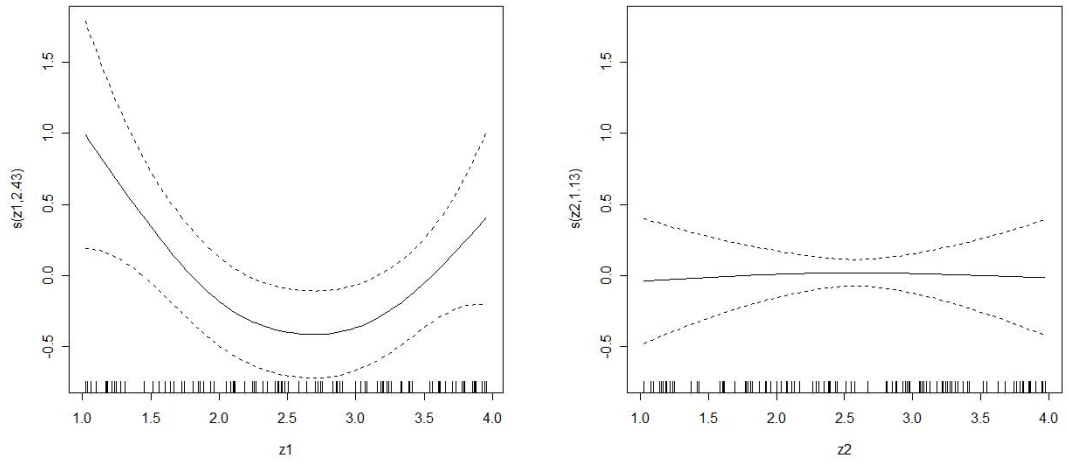
$$Y_i^{(1)} = (1 + 2 * |Z_1 - 2.5|^3) * Y_{i,eff}^{(1)} * exp(-U_i) \tag{15}$$

$$Y_i^{(2)} = (1 + 2 * |Z_1 - 2.5|^3) * Y_{i,eff}^{(2)} * exp(-U_i) \tag{16}$$

In result, Z_1 is modeled to have a cubic effect on the frontier. Z_2 is modeled as irrelevant for the frontier. The simulation is performed for $n = 100$ observations according to this scenario.[11]

The results, as illustrated in figure 1 and table 1 show that we discover the appropriate, and significant cubic effect of Z_1, while we do (correctly) not find a significant effect of Z_2. Consequently, in this additive 1-level setting, we obtain with the robust two-stage approach similar results as the conditional approach of Badin et al. (2010) with data-driven bandwidth selection. The estimated degrees of freedom are with, respectively, 2.427 and 1.134 larger than 1. This indicates that smoothing is justified by the data. The scale parameter of 0.16 indicates the presence of underdispersion in the mean-variance relation. The variability of the data is lower than expected in the binomial model. Therefore - in the robust two-stage approach - it is advisable to use the flexible quasi-binomial model instead of a binomial model.

[11] We generated a relatively small number of observations as most papers in the efficiency estimation literature also deal with small sample sizes.



(a) Effect of Z_1 (b) Effect of Z_2

Figure 1: Visualization of smooth terms in 1-level setting, n=100. The solid line presents the estimated function. Dashed lines the 95% confidence intervals. Covariate values are shown as a rug plot along the bottom of the plot.

Table 1: Summary statistics: one-level setting

Parametric coefficients

Variable	Estimate	Stand.err.
(Intercept)	1.839	0.119***

Approx. signif. of smooth terms

		Est.degrees of freedom	F-value
s(Z_1)		2.427	2.901*
s(Z_2)		1.134	0.393
Scale est.	0.161		
n	100		

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

5.2 Step 2: Explaining efficiency in a multilevel setting

We next expand the 1-level simulation to a multilevel setting. To do so, we generate $U_i \sim Expo(mean = 1/2)$, $Z_j \sim U(1, 4)$ for j=1,2 and Z_3 normally distributed between g groups; $Z_{3,k} \sim N(10, 1)$ between groups $k = 1, .., g$, with g=n/20. Output is defined by the following

production process:

$$Y_i^{(1)} = Z_{3,k} * (1 + 2 * |Z_1 - 2.5|^3) * Y_{i,eff}^{(1)} * exp(-U_i) \tag{17}$$

$$Y_i^{(2)} = Z_{3,k} * (1 + 2 * |Z_1 - 2.5|^3) * Y_{i,eff}^{(2)} * exp(-U_i). \tag{18}$$

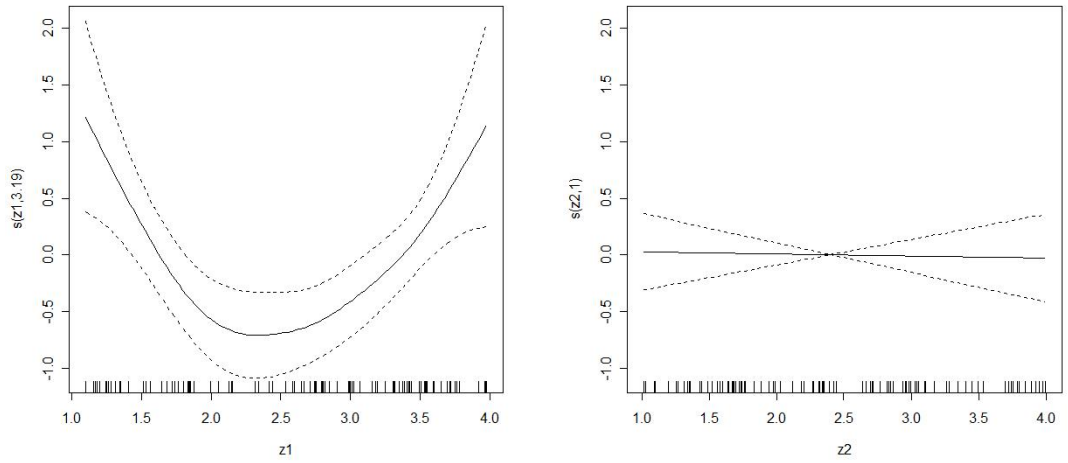
Again, as presented in Table 2 and Figure 2 we find properly the simulated significant cubic effect of Z_1, and an insignificant effect of Z_2. The estimated degrees of freedom justified by the data are equal to 1 for Z_2, which indicates that there is no need to smooth Z_2. The scale parameter of 0.13 indicates the presence of underdispersion in the mean-variance relation. It is therefore advisable to use the flexible quasi-binomial model instead of a binomial model in this setting. Consequently, in a multilevel additive setting, this simulation illustrates the usefulness of a robust two-stage approach with quasi-binomial error structure to explain efficiency.

Table 2: Summary statistics: multilevel setting

Parametric coefficients

Variable	Estimate	Stand.err.	
(Intercept)	1.869	0.1106^{***}	
Approx. signif. of smooth terms			
		Est.degrees of freedom	F-value
s(Z_1)		3.186	4.451^{***}
s(Z_2)		1.000	0.022
Scale est.	0.127		
Observations	100		

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



(a) Effect of Z_1

(b) Effect of Z_2

Figure 2: Visualization of smooth terms in 2-level setting, n=100. The solid line presents the estimated function. Dashed lines the 95% confidence intervals. Covariate values are shown as a rug plot along the bottom of the plot.

5.3 Step 3: Explaining efficiency in a multivariate and multilevel setting

This third step illustrates the usefulness of the robust two-stage approach in a multivariate and multilevel setting with the existence of categorical environmental variables. To do so, we extend the simulation in the following manner. Suppose the data are constructed in a two-level setting with groups of 20 observations ($Z_{8,k}$), that there are 5 continuous environmental variables ($Z_1 - Z_5$) - constructed as before - and 2 categorical variables ($Z_6 - Z_7$, with values 1 and 2 with equal probability). Z_1, Z_3 and Z_5 are modeled to have a cubic effect on efficiency. Z_7 is modeled to have a significant positive effect. We simulate $n = 1000$ observations.[12]

$$Y_i^{(1)} = Z_{8,k} * (1+2*|Z_1-2.5|^3) * (1+2*|Z_3-2.5|^3) * (1+2*|Z_5-2.5|^3) * Z_7 * Y_{i,eff}^{(1)} * exp(-U_i) \quad (19)$$

$$Y_i^{(2)} = Z_{8,k} * (1+2*|Z_1-2.5|^3) * (1+2*|Z_3-2.5|^3) * (1+2*|Z_5-2.5|^3) * Z_7 * Y_{i,eff}^{(2)} * exp(-U_i) \quad (20)$$

The results are presented in Table 3 and Figure 3. As was simulated, we find a significant cubic effect for Z_1, Z_3 and Z_5, a significant effect for Z_7 and an insignificant effect for Z_2, Z_4 and Z_6. The estimated degrees of freedom justified by the data is equal to 1 for Z_4. This indicates that there is no need to smooth Z_4. The scale parameter of 0.129 indicates underdispersion in the mean-variance relationship, the flexible 'quasi' approach that allows for under- and overdispersion is justified.

[12] The number of observations is increased to have a sufficient number of observations in each group.

Combining the three simulated examples clearly illustrates that the robust two-stage approach can be used to explain efficiency in multivariate and multilevel settings.

Table 3: Summary statistics: multivariate and multilevel setting

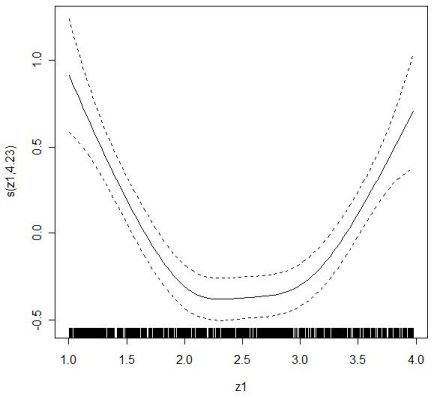
Parametric coefficients

Variable	Estimate	Stand.err.
(Intercept)	1.770	0.150^{***}
Z_6	-0.022	0.068
Z_7	0.159	0.069^{*}

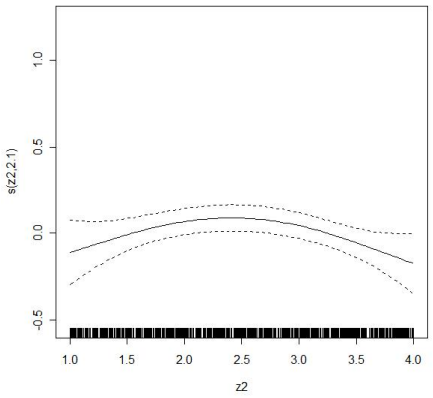
Approx. signif. of smooth terms

		e.d.f.	F-value
s(Z_1)		4.228	11.815^{***}
s(Z_2)		2.099	1.826
s(Z_3)		4.290	12.296^{***}
s(Z_4)		1.000	0.223
s(Z_5)		4.069	6.244^{***}
Scale est.	0.129		
n	1000		

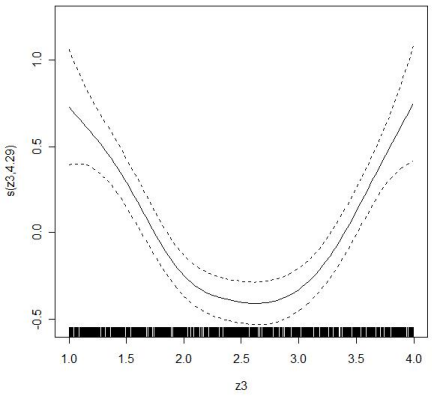
Significance levels : $*$: 5% $**$: 1% $***$: 0.1%



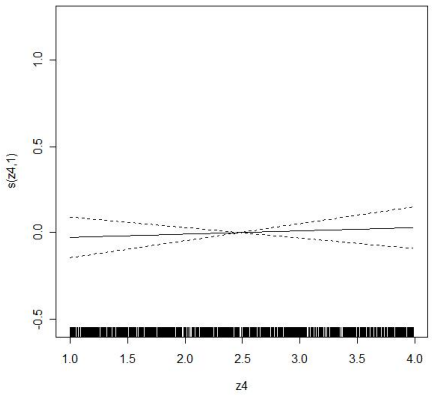
(a) Significant effect of Z_1



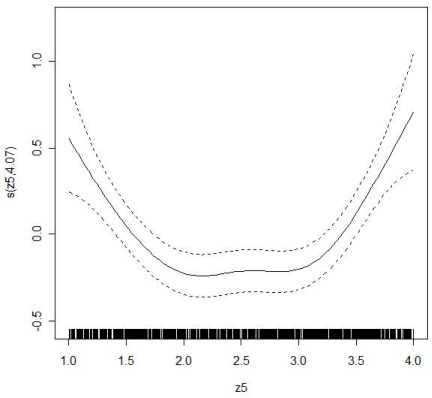
(b) Insignificant effect of Z_2



(c) Significant effect of Z_3



(d) Insignificant effect of Z_4



(e) Significant effect of Z_5

Figure 3: Visualization of smooth terms in 2-level setting, n=1000. The solid line presents the estimated function. Dashed lines the 95% confidence intervals. Covariate values are shown as a rug plot along the bottom of the plot.

6 Real data example - Influence of background on student performance

6.1 Student performance differences among Dutch provinces

In the Netherlands, there has been recently a debate about quality differences in education among the 12 Dutch provinces. In particular, in March 2009 the Dutch Ministry of Education and Culture acknowledged that the quality of education in the provinces of Groningen, Friesland and Drenthe (GFD) was lower than in other parts of the country Dijksma (2009). However, this insight is merely inspired by reports of the schools inspectorate for primary education. The school inspectorate argued that particularly the resources spend on 'care' were significantly lower in the provinces of GFD than in the other 9 Dutch provinces. Nevertheless, in our opinion, school quality is more about obtaining the highest feasible education attainments, rather than providing as much as possible 'care' for students. Additionally, when comparing the performance of students across provinces, one should control for population heterogeneity among the provinces (which was not the case in the school inspectorate report).[13]

This section attempts to examine the differences in learning capacities among the Dutch provinces. As such, we interpret 'quality' not in terms of resources for student care, but rather in terms of obtained educational attainments for given abilities and given exogenous background characteristics. We examine whether students in GFD are obtaining significantly less educational attainments, given their abilities and background, in comparison to the other provinces. In doing so, we control for random school and class specific effects in order to capture school and class heterogeneity. As we apply the outlined (semiparametric) robust two-stage model, we use a Bayesian model to structure the remaining error term.

The data are obtained from the 1999 cohort of the Voortgezet Onderwijs Cohort Leerlingen (VOCL) data. This representative data set follows students during their secondary education career. The cohort starts in the first year of secondary education with questionnaires on motivation of the students and the parents. We estimate student attainments by considering the individual test results for maths at the third year of secondary education as an output variable (i.e., results in 2002). These test scores are equivalized as to make them comparable over time.

Similar as Cherchye et al. (2010), we consider the initial test scores of students as input variables. Using the initial test scores as an input variable creates two advantages. Firstly,

[13] A complementary analysis in which we examine the impact of student counseling in the different provinces is impossible due to a lack of adequate data.

it is a proxy for the abilities of students (i.e., bright students will have higher initial test scores). Secondly, by comparing the initial test scores against the third year test scores (i.e., the output variable), we obtain a measure for what the students learn at school.[14] The initial test scores are captured by the attainment scores on math at the end of primary education (i.e., the so-called cito score; the test is taken in 1999). As such, in the setting at hand, the previously described α efficiency score denotes the probability that there is no other student with similar or lower abilities who obtains a higher math test score in the third year.

Obviously, to be fair, one should not examine the educational attainments based on the inputs (i.e., the abilities) only. The literature has rigorously indicated that the attainments of the individual student are influenced by exogenous characteristics such as (1) education of the parents, (2) ethnicity, (3) peer and track effects and (4) social segregation (see e.g. Verschelde et al. (2010) and references therein).[15] Whereas the former two variables are included at individual level in the VOCL data, the latter two variables have to be constructed.

First consider the construction of peer and track effects. The peer effects are constructed from taking the class average of the math scores (as estimated from the cito-score) in first year of secondary education. This class average of the cito-math scores presents the peer and track effects in the first year of secondary education. Following the literature, we capture peer effects additionally by including dummies for students who are living in one of the four larger Dutch cities (i.e., Amsterdam, The Hague, Rotterdam and Utrecht). The peer and track variable controls for (1) track effects and (2) peer and class effects within a track. A large literature - referenced in Hanushek et al. (2003) - has shown the difficulty to separate peer effects from other confounding class or school effects in a cross-sectional study. Therefore, no direct estimation of peer effects is undertaken. The results on peer and track effects are of an indicative nature and meant to control the setting for the mentioned unobserved track, peer and class effects.

Secondly, consider the construction of a social segregation variable. Social segregation scores are constructed by taking the school average of the parental education level. This is to correct for institutional effects that are a result of social segregation (e.g. higher involvement of parents, (self-)selection of more motivated pupils and selection of higher qualified teachers in 'richer' schools (Maaz et al., 2008). Some summary statistics are presented in Table 4.

[14] The initial test scores also indicate what students have learnt during primary education. Our analysis indicated that there are no significant differences in initial test scores across provinces.

[15] Although the VOCL data contains survey data on the motivation of the students and the parents, we did not include motivational variables because of endogeneity issues.

6.2 Results

Using the outlined robust two-stage model, we estimate four different model specifications in which alternative assumptions on the class effects and school effects, and on background influences are taken. While the first two models do not account for random class or school effects, the latter two models do. The first and third model differ from the other models as they do not control for background characteristics.

The results are presented in table 5, table 6 and figures 4. In each of the four model specifications, the scale parameter is significantly smaller than 1 (i.e., respectively 0.316, 0.284, 0.225 and 0.225), which indicates underdispersion in the mean-variance relationship. As such, the applied quasi approach is justified. We can discuss the results from 4 angles.

Firstly, if we do not control for background characteristics or random school and class effects, the results indicate provincial disparities in educational achievement, given the abilities of the students (Model 1). Pupils in Overijsel and Noord-Brabant are performing significantly better than pupils in Groningen (which is the base dummy variable). Only the province of Gelderland is consequently underperforming relatively to Groningen. Nevertheless, if we control for the previously described set of exogenous background characteristics (Model 2), only Noord-Brabant dominates significantly the province of Groningen. If in addition random class and random school effects are included in the model (Model 4), no significant regional disparities are found. This indicates that the results of the school inspectorate do not hold for secondary education if controlled for exogenous characteristics.

Secondly, we obtain from the class average of the cito-scores a clear indication for the existence of peer and track effects in the Netherlands in the beginning of secondary education. The estimated degrees of freedom for the peer and track effects amount to 3.441 and 3.414 in, respectively, Model 2 and Model 4. This suggests that smoothing of the peer effects is necessary. The 95% confidence intervals show clearly that peer effects are present in the first class of Dutch secondary education. This finding confirms previous results of De Witte (2009) who found that the first year of secondary education is crucial in shaping the dropping out decision (i.e., leaving secondary education without a diploma).

Thirdly, the estimated values of the smoothing parameter of the social segregation proxy reveal that, if not allowed for random class and school effects (i.e., Model 2), the social segregation has a significant impact on student attainment possibilities. In addition, education level of the parents significantly influences the educational attainments. Controlled for random class and school effects, we observe a linear effect of social segregation (i.e., the smoothing variable equals 1), with a significant influence of parental education. In sum, the results suggest a direct effect of family background and an indirect effect through social segregation on the

educational progress of a pupil in the first year of secondary education. This is in line with Verschelde et al. (2010), where on the basis of PISA 2006 data, the Netherlands are ranked as a country with high inequality of opportunity and significant social segregation.
Fourthly, we confirm previous literature in that individual background characteristics have a significant impact. Students of minority groups (i.e., Morocco, Turkey and Surinamese) are, given their abilities and given the education level of their parents, performing significantly less than native students.

Table 4: Descriptive statistics

Variable	Minimum	Q1	Median	Mean	Q3	Maximum
CITO math score	1.000	10.000	13.000	12.928	16.000	20.000
Class average CITO math score	3.500	11.417	13.053	12.928	14.837	20.000
Equivalized math test score	-0.842	-0.039	0.076	0.092	0.200	1.238
Highest education level parents (EDU)	2.000	4.000	4.000	4.167	5.000	7.000
School average EDU	2.000	3.885	4.154	4.167	4.423	5.500
Big city	0.000	0.000	0.000	0.061	0.000	1.000
Origin Morocco	0.000	0.000	0.000	0.017	0.000	1.000
Origin Suriname-Antilles-Aruba	0.000	0.000	0.000	0.011	0.000	1.000
Origin Turkey	0.000	0.000	0.000	0.012	0.000	1.000
Origin other country	0.000	0.000	0.000	0.085	0.000	1.000
Native pupil	0.000	1.000	1.000	0.875	1.000	1.000
Groningen	0.000			0.009		1.000
Friesland	0.000			0.040		1.000
Drenthe	0.000			0.041		1.000
Overijssel	0.000			0.030		1.000
Flevoland	0.000			0.000		1.000
Gelderland	0.000			0.118		1.000
Utrecht	0.000			0.097		1.000
Noord-holland	0.000			0.099		1.000
Zuid-holland	0.000			0.243		1.000
Zeeland	0.000			0.002		1.000
Noord-brabant	0.000			0.162		1.000
Limburg	0.000			0.157		1.000
α efficiency	0.000	0.499	0.741	0.672	0.896	1.000

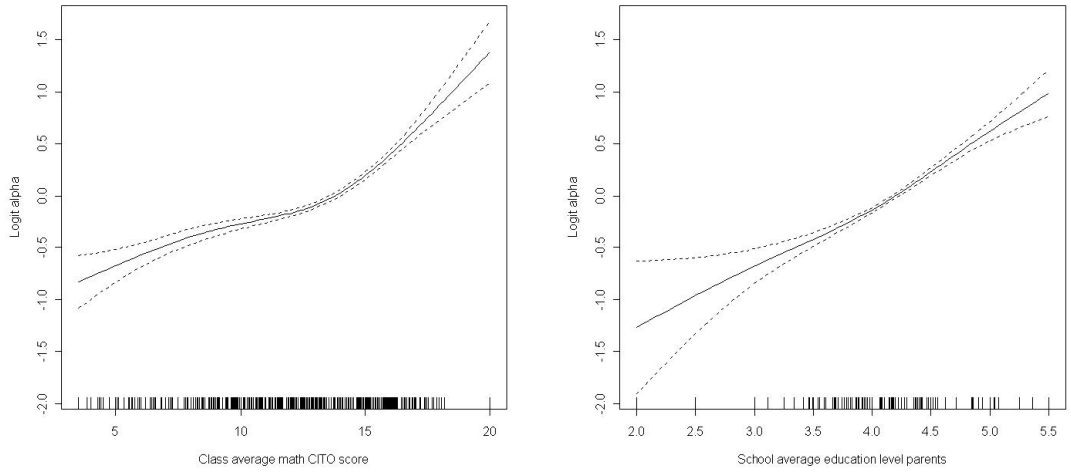
Table 5: Results: parametric coefficients

Variable	Model 1	Model 2	Model 3	Model 4
Constant	0.670***	0.243	0.509	0.352
	(0.142)	(0.149)	(0.326)	(0.239)
Friesland	-0.026	0.293	0.070	0.271
	(0.157)	(0.151)	(0.384)	(0.276)
Drente	0.271°	0.152	0.065	0.098
	(0.158)	(0.153)	(0.377)	(0.294)
Overijssel	0.424**	0.184	0.053	-0.021
	(0.164)	(0.159)	(0.385)	(0.280)
Flevoland	2.804°	1.696	1.408	1.250
	(1.586)	(1.570)	(1.454)	(1.431)
Gelderland	-0.319*	-0.145	-0.105	-0.194
	(0.147)	(0.142)	(0.356)	(0.252)
Utrecht	0.108	-0.102	0.074	-0.021
	(0.148)	(0.146)	(0.358)	(0.256)
Noord-holland	0.146	0.209	0.230	0.214
	(0.148)	(0.144)	(0.363)	(0.257)
Zuid-holland	0.020	0.073	0.154	0.139
	(0.145)	(0.140)	(0.353)	(0.248)
Zeeland	-0.012	0.477	0.181	0.528
	(0.329)	(0.315)	(0.544)	(0.402)
Noord-brabant	0.309*	0.321*	0.515	0.357
	(0.146)	(0.142)	(0.365)	(0.254)
Limburg	-0.085	0.192	0.062	0.263
	(0.146)	(0.141)	(0.398)	(0.267)
Big city		0.243***		0.066
		(0.060)		(0.083)
Maroc		-0.183°		-0.286**
		(0.100)		(0.096)
Suriname-Antilles-Aruba		-0.523***		-0.371***
		(0.117)		(0.112)
Turkey		-0.404***		-0.308**
		(0.116)		(0.110)
Other country		-0.040		-0.043
		(0.047)		(0.044)
Highest education level parents		0.100***		0.079***
		(0.014)		(0.013)
Random class effects	No	No	Yes	Yes
Random school effects	No	No	Yes	Yes
Scale est.	0.316	0.284	0.225	0.225
Observations level 1 (pupil)	8135	8135	8135	8135
Observations level 2 (class)		598		598
Observations level 3 (school)		106		106
Significance levels : ∘ : 10% :	∗ : 5%	∗∗ : 1%	∗ ∗ ∗ : 0.1%	

Table 6: Results: approximate significance of smooth terms

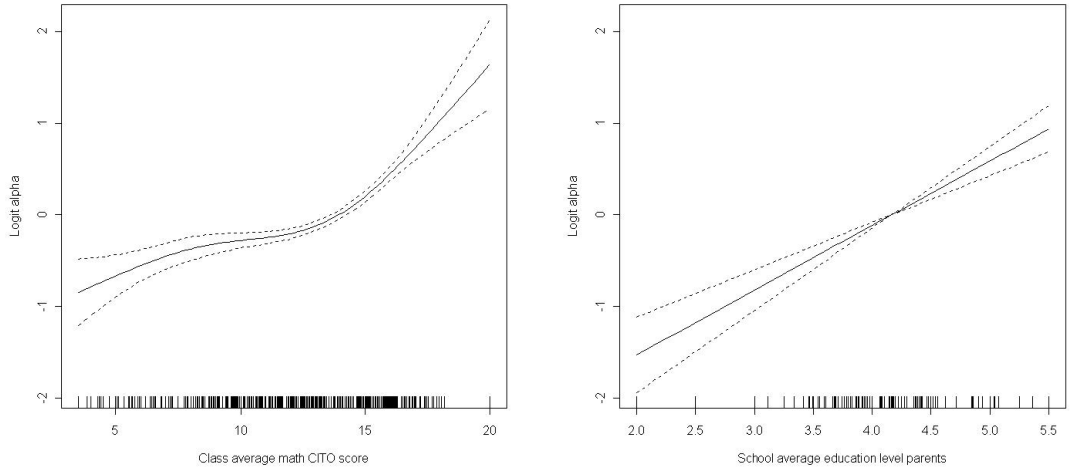
Variable		Model 1	Model 2	Model 3	Model 4
s(class average CITO math score)	e.d.f.		3.441		3.414
	F-value		88.93***		55.21***
s(school average education level parents)	e.d.f.		2.405		1.000
	F-value		67.49***		39.53***
Random class effects		No	No	Yes	Yes
Random school effects		No	No	Yes	Yes
Scale est.		0.316	0.284	0.225	0.225
Observations level 1 (pupil)		8135	8135	8135	8135
Observations level 2 (class)			598		598
Observations level 3 (school)			106		106

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



(a) Model 2: effect of class average CITO math score

(b) Model 2: effect of school average education level parents



(c) Model 4: effect of class average CITO math score

(d) Model 4: effect of school average education level parents

Figure 4: Results: visualization of smooth terms. Solid line represents the estimated function. Dashed lines the 95% confidence intervals. Covariate values are shown as a rug plot along the bottom of the plot.

7 Conclusion

Many applications are characterized by a clustered, hierarchical data generating process. The traditional non-parametric efficiency analysis literature does not include a smooth way of estimating and explaining efficiency in a highly multi-level context. Indeed, the traditional

two-stage approach (in which deterministic Data Envelopment Analysis estimates are parametrically regressed on background characteristics) has been heavily criticized by Simar and Wilson (2007). The conditional efficiency approach (Daraio and Simar, 2005) suffers from an inconvenient slow bandwidth computation if too many variables are included.

This paper suggested an alternative robust approach which is specially tailored for highly multi-level frameworks. The proposal estimates in a first stage the robust α-efficiency estimator of Daraio and Simar (2007a) and Aragon et al. (2006). Thanks to the attractive features of this estimator, the proposed framework carefully avoids the Simar and Wilson (2007) critique. In a second phase, the α-efficiency estimators are regressed on the exogenous background characteristics by a Generalized Additive Mixed Model (GAMM). Besides being extremely flexible, and requiring only few assumptions, the GAMM approach allows us to estimate the over- or underdispersion in the error structure. The appropriate working of the technique is illustrated on simulated data.

Finally, the proposed model is applied on a rich sample of Dutch educational data. In particular, we examined a claim of the Dutch Ministry of Education in that the quality of education is lower in three out of the twelve provinces. Contrary to the Ministry of Education, we analyzed the educational attainments of students while controlling for a broad set of background characteristics. Our results suggest that, if properly controlled for the exogenous environment, there are no differences in educational attainments among the Dutch provinces. As a side effect of our estimations, we find strong social segregation among schools and classes.

This paper opens several avenues for further research. Firstly, the procedure could be further developed by making it fully non-parametric. Therefore, the GAMM model could be altered. Secondly, as it stands now, the procedure suffers from a separability approach. Therefore, in a next step, the α-efficiency scores could be altered such that they account for the background characteristics. Finally, many real life applications could be developed by using this technique.[16] Indeed, finally the literature on efficiency estimations can estimate panel models with fixed and random effects, as it can estimate highly multilevel estimations.

8 Appendix

8.1 Numerical illustration of the robust two-stage approach

Simulation 1: consistency and robustness of α To illustrate the robust features of the robust two-stage approach, we follow the example of Simar and Wilson (2008, p. 488-491). The production set ψ is bounded above by the concave frontier $y_{eff} = g(x) = (2x - x^2)^{1/2}$. The probability density function $f(x, y)$ is given by (21). $f(x, y)$ is uniform over ψ (see

[16] Therefore, the R code is available from the authors upon simple request.

fig (5(a)). The marginal density of x is given by (22). The marginal distribution function of x is given by (24). The true α efficiency is given by (26).

$$f(x,y) = \begin{cases} \frac{4}{\pi} \forall x \in [0,1], y \in [0, g(x)], \\ 0 \text{ otherwise.} \end{cases} \tag{21}$$

$$f(x) = \int_{g(x)}^{0} f(x,y)dy = \frac{4}{\pi}(2x - x^2)^{1/2} \tag{22}$$

$$F_X(x) = Prob(X \leq x) \tag{23}$$

$$= \begin{cases} 1 & \forall x > 1; \\ 4\pi^{-1}[\frac{x-1}{2}((2x - x^2)^2)^{1/2} + \frac{1}{2}sin^{-1}x_0 - 1) + \frac{\pi}{4} & \forall x \in [0,1]; \\ 0 & \forall x < 0 \end{cases} \tag{24}$$

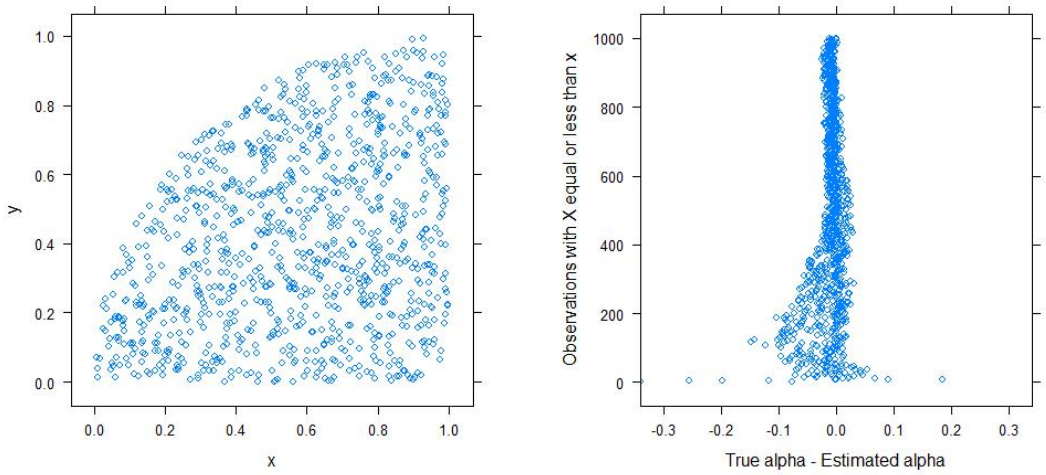
$$F_{Y|X}(y\,|X \leq x) = Prob(Y \leq y\,|X \leq x) \tag{25}$$

$$= \begin{cases} 1 & \forall y \geq g(x); \\ F_X(x)^{-1}4\pi^{-1}\left\{(x-1)y + \frac{1}{2}[y(1-y^2)^{1/2} + sin^{-1}(y)]\right\} & \forall y \in [0, g(x)]; \\ 0 & \forall y < 0 \end{cases} \tag{26}$$

Table 7 illustrates that $\hat{\alpha}$ is a biased, but consistent estimator of α. The effect of outliers is small and goes to zero when the sample size goes to infinity. Figure (5(b) illustrates that, by construction, in large sample, there are still units where $\hat{\alpha}$ is severely imprecise. However, when $n \to \infty$, the proportion of units (x_i, y_i) with few units j with $X_j \leq x_i$ in the whole sample tends to 0.

Table 7: The bias of $\hat{\alpha}$

Observations	mean($\alpha - \hat{\alpha}$)	with 3 outliers
for n=25 and m=1000	-0.059	-0.032
for n=100 and m=1000	-0.020	-0.014
for n=1000 and and m=1000	-0.003	-0.002



(a) The setting, 1000 observations

(b) Consistency of alpha efficiency estimation

Figure 5: Numerical illustration

8.2 An overview of Generalized Additive (Mixed) Modelling

Generalized Linear Model (GLM) In the seminal work of Nelder and Wedderburn (1972), Generalized linear models are proposed (1) to allow for response distributions other than normal and (2) to introduce non-linearity in the model structure. Formally, a GLM model is represented as

$$E(Y) \equiv \mu,\ g(\mu) = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + \dots \tag{27}$$

where g denotes a known monotonic link function, X the explanatory variables and β the parameters to be estimated.

Generalized Additive Model (GAM) To avoid parametric assumptions on the unknown relationship between the response variable and the explanatory variables, Generalized additive Models (GAMs) as popularized by Hastie and Tibshirani (1990) can be used. A GAM is a generalized linear model where the linear predictor is specified as a sum of smooth functions of some or all the covariates (Wood and Augustin, 2002).

$$E(Y) \equiv \mu,\ g(\mu) = \beta_0+\beta_1(x_1)+\beta_2(x_2)+\ldots+\beta_p(x_p)+s_1(x_{p+1})+s_2(x_{p+2})+\ldots+s_q(x_{p+q}) \tag{28}$$

here X present the fixed effects variables with p variables modeled parametrically and q variables modeled nonparametrically. For this, q unknown smooth functions s_i, with $i = 1, \dots, q$

are defined.

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 as advantage that multiple smooth terms can be included. The largest disadvantage is 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. However, the high computational cost of the Gu and Wahba (1991) approach is an important practical barrier. A penalized regression spline approach as proposed in Eilers and Marx (1996), Marx and Eilers (1998), Wahba (1980) and Wahba (1990) is a computationally efficient approach to estimate a GAM with integrated model selection.

To represent smooth functions, known basis functions are used. Consider a set of basis functions $\{b_k(x) : k = 1, \ldots, m\}$. Commonly used examples are the polynomial basis function and the cubic basis function. A k-dimensional polynomial basis is defined as: $b_1(x) = 1, b_2(x) = x, b_3(x) = x^2, \ldots, b_k = x^{(k-1)}$(Wood and Augustin, 2002). For a set of - usually evenly placed - points in the range of x (='knots') $\left\{x_j^* : k = 1, \ldots, m\right\}$, the cubic basis can be defined as $b_k(x) = |x - x_j^*|^3$, for $j = 1, \ldots, m$, with $b_{m+1} = 1$,$b_{m+2}(x) = x$ (Wood and Augustin, 2002). Then $s(x)$ can be represented as the sum of basis functions, multiplied by the respective basis parameters α_j, with $j = 1, \ldots, m$.

$$s(x) = \sum_{j=1}^{m} \alpha_j \beta_j(x) \tag{29}$$

where α_j are m unknown coefficients.

Spline smoothing has better approximation theory and higher numerical stability with cubic basic functions than with polynomial basis functions (Wood and Augustin, 2002).

The penalized regression spline approach introduces a penalty λ_j for wiggliness of smooth function s_k, with $k = 1, \ldots, q$ to avoid overfitting of the smooth functions as is the case in MLE of GAM models (Wood and Augustin, 2002). The deviance function to be minimized is defined as follows

$$D(\mu) - \sum_k \lambda_k \times [\text{wiggliness of } s_k] \tag{30}$$

here $\mu = E(Y)$ and deviance function D is $2(l(Y) - l(\mu))$, with l the log likelihood.

The wiggliness can be defined as

$$\int f_k''(x)^2 dx = \boldsymbol{\beta}_k \mathbf{S}_k \boldsymbol{\beta}_k \tag{31}$$

with $\mathbf{S}_k$ a matrix of known basis coefficients. The minimization problem can be defined as:

$$\text{minimize } D(\mu) + \sum_k \lambda_k \boldsymbol{\beta}_k \mathbf{S}_k \boldsymbol{\beta}_k \text{ w.r.t.} \boldsymbol{\beta}. \tag{32}$$

Estimation of $\boldsymbol{\beta}$ and λ can be done by penalized iteratively re-weighted least squares (P-IRLS), using Generalized Cross Validation (GCV) or an UnBiased Risk Estimator (UBRE) as thoroughly discussed in Wood and Augustin (2002), Wood (2004) and Wood (2006).

Generalized Additive Mixed Model (GAMM) To allow for overdispersed and correlated data in a clustered, hierarchical design, extention of the generalized model to the mixed model setting is needed (Lin and Zhang, 1999). Since the work of Breslow and Clayton (1993), GLMs are extended to include random effects. To avoid the hard parametric mean assumption in a GLMM, Generalized Additive Mixed Models are proposed (Lin and Zhang, 1999). A GAMM with identity link can be defined as

$$E(Y) \equiv \mu,\ g(\mu) = \beta_0+\beta_1(x_1)+\beta_2(x_2)+\ldots+\beta_p(x_p)+s_1(x_{p+1})+s_2(x_{p+2})+\ldots+s_q(x_{p+q})+Z_i\mathbf{b}+\epsilon_i \tag{33}$$

where Y denotes the response variable, β_f, with $f = 1, \ldots, p$ the fixed parameters, s_k, with $k = 1, \ldots, q$ the smooth functions, Z_i a row of the random effects model and $b \sim N(0, \phi_\theta)$ is a random effects coefficient, with unknown positive definite covariance matrix ϕ_θ, with parameter θ and $\epsilon \sim N(0, \Lambda)$ the residual error vector with covariance matrix Λ (Wood, 2006).

It can be shown that each penalized regression smoother can be included in a mixed model (Wood, 2006). The smooths are treated to have a fixed -unpenalized - component, which can be absorbed in $X_i\boldsymbol{\beta}$ and a random effects -penalized - component, which can be absorbed in $Z_i\mathbf{b}$. As for a GLMM , penalized quasi-likelihood approaches as proposed by Breslow and Clayton (1993) can be used to estimate a GAMM.

This involves an iterative minimization process of an approximation of the understanding likelihood function

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) \propto |\phi_{\boldsymbol{\theta}}|^{-1/2} \int exp\left(l(\boldsymbol{\beta}, \mathbf{b}) - \frac{1}{2}\mathbf{b}^T\phi_{\boldsymbol{\theta}}^{-1}\mathbf{b}\right) d\mathbf{b} \tag{34}$$

where $l(\boldsymbol{\beta}, \mathbf{b})$ is the log likelihood of the GLM that would results from treating both $\boldsymbol{\beta}$ and $\mathbf{b}$ as fixed effects (Wood, 2006).

As discussed in Wood and Augustin (2002), Bayesian approximation can be used to construct confidence intervals for the smooth functions.

References

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

Aragon, Y., Daouia, A., Thomas-agnan, C., 2006. Efficiency measurement: a nonparametric approach. Annales d'économie et de statistique (82), 27.

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.

Banker, R., Natarajan, R., 2008. Evaluating contextual variables affecting productivity using data envelopment analysis. Operations Research 56, 48–58.

Breslow, N., Clayton, D., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88, 9–25.

Cazals, C., Florens, J. P., Simar, L., Jan. 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.

Cherchye, L., De Witte, K., Ooghe, E., 2010. Equity and efficiency in private and public education: a nonparametric comparison. European Journal of Operational Research 202, 563–573.

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., 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.

De Witte, K., 2009. Dropout from secondary education: all's well that begins well. Top Institute for Evidence Based Education Research - Discussion paper 09.03.

De Witte, K., Kortelainen, M., 2008. Blaming the exogenous environment? conditional efficiency estimation with continuous and discrete environmental variables. Center for Economic Studies - Discussion papers 0833.

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

Deprins, D., Simar, L., Tulkens, H., 1984. The performance of public enterprises - concepts and measurement. Amsterdam, North-Holland, Ch. Measuring labor-efficiency in post offices, pp. 243–267.

Dijksma, S., 2009. Kwaliteit onderwijs noordelijke provincies. De voorzitter van de Tweede Kamer der Staten Generaal, 1–2.

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

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

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.

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.

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

Hjalmarsson, L., Kumbhakar, S., Heshmati, A., 1996. DEA, DFA and SFA: a comparison. Journal of Productivity Analysis 7 (2), 303–327.

Lin, X., Zhang, D., 1999. Inference in generalized additive models using smoothing splines. Journal of the Royal Statistical Society B 61, 381–400.

Maaz, K., Trautwein, U., Lüdtke, O., Baumert, J., 2008. Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational outcomes. Child Development Perspectives 2 (2), 99–106.

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

McDonald, J., 2009. Using least squares and tobit in second stage dea efficiency analysis. European Journal Of Operational Research 197, 792–798.

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

Nelder, J., Wedderburn, R., 1972. Generalized linear models. Journal of the Royal Statistical Society, Series A 135, 370–384.

Papke, L. E., Wooldridge, J. M., 1996. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. Journal Of Applied Econometrics 11 (6), 619–632.

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.

Raudenbush, S. W., Bryk, A. S., 1986. A hierarchical model for studying school effects. Sociology of Education 59 (1), 1–17.

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

Simar, L., Wilson, P., 2008. Statistical inference in nonparametric frontier models: recent developments and perspectives, in Fried, H., and Lovell, C.A.K., and 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., Zelenyuk, V., 2007. Statistical inference for aggregates of farrell-type efficiencies. Journal Of Applied Econometrics 22 (7), 1367–1394.

Venables, W. N., Dichmont, C. M., 2004. Glms, gams and glmms: an overview of theory for applications in fisheries research. Fisheries Research 70 (2-3), 319–337.

Verschelde, M., Hindriks, J., Rayp, G., Schoors, K., 2010. Ability tracking and equality of opportunity in schooling: evidence from belgium.

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

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

Wood, S., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99, 673–686.

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.

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