Blaming the exogenous environment? Conditional efficiency estimation with continuous and discrete environmental variables*

Kristof De Witte
Centre for Economic Studies
University of Leuven (KU Leuven)
Naamsestraat 69, 3000 Leuven, Belgium
kristof.dewitte@econ.kuleuven.be

Mika Kortelainen
Aston Business School
Aston University, Aston Triangle
Birmingham, B4 7ET, UK
m.kortelainen@aston.ac.uk

December 7, 2008
Preliminary version

Abstract

This paper proposes a fully nonparametric framework to estimate relative efficiency of entities while accounting for a mixed set of continuous and discrete (both ordered and unordered) exogenous variables. Using robust partial frontier techniques, the probabilistic and conditional characterization of the production process, as well as insights from the recent developments in nonparametric econometrics, we present a generalized approach for conditional efficiency measurement. To do so, we utilize a tailored mixed kernel function with a data-driven bandwidth selection. So far only descriptive analysis for studying the effect of heterogeneity in conditional efficiency estimation has been suggested. We show how to use and interpret nonparametric bootstrap-based significance tests in a generalized conditional efficiency framework. This allows us to study statistical significance of continuous and discrete environmental variables. The proposed approach is illustrated by a sample of British pupils from the OECD Pisa data set. The results show that several exogenous discrete factors have a significant effect on the educational process.

Keywords: Nonparametric estimation, Conditional efficiency measures, Environmental factors, Generalized kernel function, Education

JEL-classification: C14, C61, I21

*We would like to thank the participants of the Seminar on Efficiency and Productivity Analysis at Aston University for valuable comments.

1 Introduction

The traditional nonparametric procedures to estimate efficiency (such as the non-convex Free Disposal Hull (FDH; Deprins *et al.*, 1984) and the convex Data Envelopment Analysis (DEA; Charnes *et al.*, 1978) have recently been directed towards the incorporation of exogenous environmental variables. Indeed, efficiency estimations which do not account for the operational environment have only a limited value. If, for example, the efficiency of the educational system is assessed, it is useless to compare schools located in 'good' neighborhoods (e.g. measured by the highest degree of the mother, income of the parents, native language) with schools located in less advantageous areas. Thus, if the evaluated observations are affected by external, exogenous factors, performance analysis should control for this heterogeneity.

The literature counts various approaches to incorporate the exogenous environment in nonparametric efficiency analysis (for an overview see Fried *et al.*, 2008). The first family of models uses a *one-stage* approach (e.g. Banker and Morey, 1986a, 1986b; Ruggiero, 1996), where environmental factors are considered as free disposable inputs and/or outputs which are used in the estimation of the production possibility set, but treated as non-controllable (or non-discretionary) variables. Essential drawbacks of this approach are that (1) the researcher has to choose *a priori* whether to model the environmental variable as an input or as an output, and that (2) the environmental variable is required to be free disposal (monotone) in the production process (and possibly also convex if DEA is used).[1] Although several variants have been developed (e.g. Färe *et al.*, 1989; Ferrier and Lovell, 1990), they also suffer from problem (1) and are only suitable for continuous variables.

The second family of models is based on a so-called *frontier separation* (or *metafrontier*) approach (e.g. Charnes *et al.*, 1981; Thanassoulis and Portela, 2002; Battese *et al.*, 2004; De Witte and Marques, 2008a). The basic idea behind the frontier separation approach is to group evaluated units according to some criteria and then perform separate efficiency assessments for different groups (or different values of environmental variables). However, this approach (1) can only be applied to categorical environmental variables, and (2) in practice it is not possible to include several environmental factors (at least if variables have many classes) because otherwise groups can be very small. In addition, (3) comparison and statistical testing of efficiency differences between more than two categories seems to be challenging.

The third family of models is based on a *two-stage* approach (e.g. Ray, 1991; Simar and

[1] Note that these conditions have to be satisfied for both continuous and discrete environmental variables. For catecorigal variables this means that the researcher needs to order beforehand the values of the variable from the least to most detrimental effect upon efficiency. This is very restrictive and in practice one cannot use variables that are unordered (e.g. school). Moreover, the approach is not meaningful in applications with several environmental variables (see Ruggiero, 1998).

Wilson, 2007; Park *et al.*, 2008). Environmental factors are not included in the first stage of the efficiency estimation, but only in a second stage regression model where the efficiency scores are explained by the environmental variables. Also three- and four-stage approaches (see Ruggiero, 1998; Fried *et al.*, 1999, 2002) have been proposed. Different multi-stage models avoid the above problems and make it possible to include both continuous and discrete variables. However, the multi-stage approaches assume (implicitly) a separability condition in that the operational environment would not influence the input or output levels, but only efficiency. Obviously, in many applications the exogenous variables (e.g. the neighborhood and mother tongue) do influence the observed input use (e.g. teaching hours) and output levels (e.g. test scores) of the observations. In this sense, there is no separability between the inputs and outputs on the one hand, and the exogenous variables on the other hand.

The fourth and more novel approach for including environmental factors is based on a probabilistic formulation of the production process. It incorporates the operational environment by conditioning on the exogenous characteristics (Cazals *et al.*, 2002; Daraio and Simar, 2005, 2007a). This so-called *conditional efficiency* approach generalizes previous models by avoiding the separability condition and by not requiring any specification on the direction of influence of environmental variables. In addition, it allows one to include several environmental variables and to examine the effect (favorable or unfavorable) of them. As the conditional efficiency approach avoids the problems of the other models, it seems to be the most promising method to introduce external environmental factors into nonparametric frontier models. Therefore, the remainder of this paper concentrates on this approach.

Cazals *et al.* (2002) outlined the idea on how to incorporate exogenous variables in the non-convex nonparametric model. Daraio and Simar (2005, 2007a) expanded their approach to a more general multivariate (continuous) setup and presented a practical methodology to evaluate the estimators. Later, extension to convex nonparametric models was proposed (Daraio and Simar, 2007b) and also a significant amount of work has been done to prove the consistency and the asymptotic properties of different conditional efficiency estimators (Cazals *et al.,* 2002; Jeong *et al.*, 2008). As the merits of the approach are large (in particular avoiding the separability condition) it is increasingly used in several research questions. Previous applications include the productivity of universities (Bonaccorsi *et al.*, 2006, 2007a, 2007b; Bonaccorsi and Daraio, 2008), efficiency in the water sector (De Witte and Dijkgraaf, 2008; De Witte and Marques, 2008b, De Witte and Saal, 2008), performance of mutual funds (Daraio and Simar, 2005, 2006; Daouia and Simar, 2007; Jeong *et al.*, 2008; Badin *et al.*, 2008) and banks (Blass Staub and da Silva e Souza, 2007), efficiency of post offices (Cazals *et al.*, 2008), knowledge spillover and regional innovation performance (Bonaccorsi and Daraio, 2007c; Broekel, 2008; Broekel and Meder, 2008) and primary education (Cherchye *et al.,* 2007).

Nevertheless, some intricate issues remain. As the conditional efficiency approach relies on the estimation of nonparametric kernel functions to select the appropriate reference partners, it heavily relies on the bandwidths. The original article of Daraio and Simar (2005) considered the cross-validation k-nearest neighbor technique for estimating the bandwidth. However, besides being nonoptimal in finite samples, this bandwidth approach does not take into account the influence of the exogenous variable on the production process. As such, although the conditional efficiency estimates avoid the separability condition, its bandwidth relied on it. Recently, Badin *et al.* (2008) suggested an alternative (i.e. data-driven) approach to select the optimal bandwidth. This approach accounts for the input and output variables while selecting the bandwidth. Moreover, following Hall *et al.* (2004), a data-driven procedure can help to identify external variables that have no influence on the production process.

The current paper contributes to the literature by focussing on three additional issues, which are very relevant in most empirical applications. Firstly, it considers the inclusion of both discrete and continuous variables in the conditional efficiency framework. The conditional models used in previous studies have been designed for continuous environmental variables only.[2] However, in interesting real-life applications the exogenous variables are both continuous and discrete. This paper shows how to adapt the nonparametric conditional efficiency measures to include mixed (i.e. both continuous and discrete) exogenous variables by specifying an appropriate kernel function which smooths the mixed variables. In doing so, we propose a procedure to estimate (data-driven) kernel bandwidths both for continuous and discrete variables (adapted from Hall *et al.*, 2004). By estimating an observation and variable specific bandwidths, our approach is able to estimate for every observation the efficiency relative to a sufficiently large reference group of similar units (i.e. units with a large probability of being similar). As such, the approach is also superior to the frontier separation approach which dramatically reduces the number of reference units when the number of groups is large.

Secondly, thanks to our specific kernel estimation, our approach can include several ordered and/or unordered categorical variables along with continuous environmental variables even in relatively small samples. We know from previous research (Cazals *et al.*, 2002; Jeong *et al.*, 2008) that the convergence rate of conditional efficiency estimators decreases when the number of continuous environmental variables increases. The typical curse of dimensionality in nonparametric models is deteriorated in the conditional efficiency models due to the smoothing on the exogenous variables. Owing to this problem, one cannot include many continuous environmental variables in small samples. However, we know from nonparametric econometrics and statistics that discrete variables with compact support are not sensitive to this dimensionality problem (see e.g. Li and Racine, 2007). We argue and prove that this is

[2] In some applications, it might be justified to use continuous kernels for ordered dicrete variables with many categories, since those variables are close to be continuous. Instead, the values of unordered discrete variables have no natural order, and thus cannot be modelled analogously with continuous variables.

also the case in our conditional efficiency framework. In particular, we prove that the convergence rate does not depend on the number of discrete variables. This is very relevant for applied research, because it allows one to include a large number of discrete environmental variables in conditional efficiency measures without deteriorating accuracy of estimation.

Thirdly, we present a framework to test nonparametrically the significance of the exogenous variables. We note that, so far, only descriptive analysis for studying the effect of the environmental variables in conditional efficiency estimation has been suggested (Daraio and Simar, 2005). This is in contrast to the two-stage semiparametric approach of Simar and Wilson (2007), which allows one to evaluate the significance of explanatory variables in a truncated regression by the use of bootstrapping techniques. We extend the Daraio and Simar toolbox for visualizing the effects of the continuous exogenous variables to a generalized nonparametric setting which allows both visualizing and statistical inference of continuous and discrete environmental variables. For the significance testing, we use recently developed nonparametric boostrap-based procedures.

Thanks to our contributions, the nonparametric setup starts approaching very well the benefits of a parametric model (i.e. multivariate analysis with continuous and discrete factors and with well established statistical inference), but without facing the major drawback of a parametric model (i.e. selecting *a priori* a functional form of the production process). To show potentiality of the approach, we demonstrate it by a relevant research question. In particular, the inclusion of both discrete and continuous variables in the conditional efficiency estimates is illustrated by assessing the efficiency of a random sample of British 15 years old pupils. We use the Pisa data set (Program for International Student Assessment) to estimate the performance of pupils while accounting for a broad range of unordered (e.g. mother tongue, possession of own room), ordered (highest degree of mother and father) and continuous environmental variables (school size or teacher-student ratio). Including both discrete and continuous factors in the nonparametric model allows for a rich and solid analysis. Obviously, our approach is not limited to educational performance assessment but could be implemented in about all known applications.

The remainder of the paper unfolds as follows. The next section discusses the probabilistic formulation of the production process and describes the conditional efficiency approach. Section 3 presents the generalized kernel estimation, its appropriate bandwidth selection and shows the procedure for testing the significance of environmental variables. Section 4 applies the insights to the Pisa data set. Finally, we present the conclusions.

2 Conditional order-m efficiency estimators

2.1 Probabilistic formulation and order-m

Nonparametric efficiency measures are based on micro-economic production theory and estimation methods that do not require any functional form assumptions. In this framework it is typical to consider a production technology where production units are characterized by a set of inputs x ($x \in \mathbb{R}_+^p$) and outputs y ($y \in \mathbb{R}_+^q$). The production technology is the set of of all feasible input-output combinations: $\Psi = \left\{(x,y) \in \mathbb{R}_+^{p+q} \mid x \text{ can produce } y\right\}$. Various efficiency measures can be defined using the set Ψ. For example, the traditional Farrell (1957) output-oriented technical efficiency measure is usually defined as:

$$\lambda(x,y) = \sup\left\{\lambda \mid (x,\lambda y) \in \Psi\right\}, \tag{1}$$

where the output efficiency measure $\lambda(x,y) \geq 1$ is the proportionate increase of outputs, which the unit operating at level (x,y) should attain to be considered as being efficient (i.e. $\lambda(x,y) = 1$).

Obviously, in practice the set Ψ and the efficiency measures are unknown and have to be estimated from a random sample of production units denoted by $\chi_n = \{(x_i,y_i) \mid i = 1,...,n\}$.[3] To make the estimation operational, we need to make some assumptions regarding the production possibility set Ψ. One usual assumption is the free disposability of inputs and outputs, defined as: $\forall (x,y) \in \Psi$, if $\widetilde{x} \geq x$ and $\widetilde{y} \leq y$ then $(\widetilde{x},\widetilde{y}) \in \Psi$. This assumption, which can be easily defended in most applications, is typically required in the nonparametric efficiency framework. For example, the non-convex Free Disposal Hull model (FDH, Deprins *et al.*, 1984) relies only it, while the convex Data Envelopment Analysis (DEA, Charnes *et al.*, 1978) estimators require it along with an additional convexity assumption. In the remainder of this paper, we will concentrate on the FDH model as its free disposability assumption can be easily defended, whereas the convexity assumption is more intricate. The FDH model estimates the production possibility set as:

$$\hat{\Psi}_{FDH} = \left\{(x,y) \in \mathbb{R}_+^{p+q} \mid y \leq y_i, x \geq x_i, (x_i,y_i) \in \chi_n\right\}. \tag{2}$$

By interpretation, FDH estimates the set Ψ using best practice observations that are defined as undominated units which produce with a given input vector x the highest output vector y (i.e. an output-orientation), or alternatively, which are able to produce a set of outputs y with the smallest set of inputs x (i.e. an input-orientation). Note that FDH estimator for the Farrell output-oriented efficiency score is then obtained by replacing Ψ with $\hat{\Psi}$ in the equation (1).

[3] To clarify presentation, we denote the observed sample from which the efficiency scores are estimated by lowercase letters (x_i,y_i) whereas uppercase letters (X,Y) denote the unknown (and thus random) variables which can take any value.

Traditionally, the production process and different efficiency measures have been presented using the production possibility set $\hat{\Psi}$ as illustrated above. Recently, Cazals *et al.* (2002) described the production process using an equivalent probabilistic formulation, which provides an alternative way of describing the nonparametric FDH estimators. The probabilistic formulation is also useful in presenting a robust version of FDH and in introducing environmental factors in the nonparametric framework (see below). The idea behind this probabilistic formulation is to examine the probability that an evaluated observation (x, y) is dominated using the joint probability function:

$$H_{XY}(x, y) = \Pr(X \leq x, Y \geq y). \tag{3}$$

It is worth emphasizing that $H_{XY}(x, y)$ is not a standard distribution function, because for the outputs y the survival form is used, not the cumulative form like for the inputs x. In line with the idea of FDH, $H_{XY}(x, y)$ gives the probability that a unit, operating at input-output levels (x, y), is dominated. The joint probability function can be further decomposed as (remark: we only present the output-orientation, for the input-orientation see Cazals *et al.*, 2002):

$$\begin{aligned} H_{XY}(x, y) &= \Pr(Y \geq y \mid X \leq x)\ \Pr(X \leq x) \\ &= S_{Y|X}(Y \geq y \mid X \leq x) F_X(X \leq x) \\ &= S_Y(y \mid x)\ F_X(x) \qquad \textit{(in shorthand notation)} \end{aligned} \tag{4}$$

where $S_Y(y \mid x)$ denotes the conditional survivor function of Y and $F_X(x)$ the cumulative distribution function of X. Now it easy to show that if Ψ is free disposal (as assumed above), the upper boundary of the support of $S_Y(y \mid x)$ defines the traditional Farrell output-oriented technical efficiency measure:

$$\lambda(x, y) = \sup\{\lambda \mid S_Y(\lambda y \mid x) > 0\} = \sup\{\lambda \mid H_{XY}(x, \lambda y) > 0\}. \tag{5}$$

This alternative presentation of the output-oriented efficiency score can be interpreted as the proportionate increase in outputs required for the evaluated unit to have zero probability of being dominated at the given input level.

To estimate efficiency scores using the probabilistic formulation, one needs to first substitute the empirical distribution function $\widehat{H}_{XY,n}(x, y)$ for $H_{XY}(x, y)$ and $\widehat{S}_{Y,n}(y \mid x)$ for $S_Y(y \mid x)$, correspondingly. These empirical analogs are given by:

$$\widehat{H}_{XY,n}(x, y) = \frac{1}{n} \sum_{i=1}^{n} I\left(x_i \leq x, y_i \geq y\right) \tag{6}$$

and

$$\widehat{S}_{Y,n}(y \mid x) = \frac{\widehat{H}_{XY,n}(x, y)}{\widehat{F}_{X,n}(x)} = \frac{\widehat{H}_{XY,n}(x, y)}{\widehat{H}_{XY,n}(x, 0)}, \tag{7}$$

where $I(\cdot)$ is an indicator function. Using the plug-in principle, the FDH estimator for the output-oriented efficiency score can be then obtained as $\widehat{\lambda}_{FDH}(x,y) = \sup\left\{\lambda \mid \widehat{S}_{Y,n}(\lambda y \mid x) > 0\right\}$.

It should be noted that the traditional FDH estimator $\widehat{\lambda}_{FDH}(x,y)$ has two major drawbacks: (1) it is deterministic and (2) it does not account for the operational environment. Here we discuss the first issue, while the second one is treated in the next subsection. The deterministic nature of the FDH estimator arises from the assumption that all observations constitute the production set: $Prob((x,y)\ \epsilon\ \Psi) = 1$. As such, the nonparametric technique is sensitive to outlying and atypical observations as these could heavily influence the upper boundary of the support of $\widehat{S}_{Y,n}(y \mid x)$. Therefore, Cazals *et al.* (2002) suggested to consider instead of the maximum output production for a given input (which could be influenced by atypical observations), the expected value of m random variables $Y_i, i = 1, ..., m$ generated by the conditional q-variate distribution function $S_Y(y \mid x)$. Thus, instead of considering the full frontier, the idea is to draw a partial frontier depending on a random set of m variables which consume maximally x resources. Taking the expectation of this less extreme benchmark, we obtain the order-m efficiency measure $\lambda_m(x,y)$. If a unit is on average performing superior than its m randomly drawn (with $X \leq x$) reference units, it obtains a 'super-efficiency' score (i.e. an output-efficiency score of $\lambda_m(x,y) < 1$) which is impossible in the traditional framework where by construction $\lambda(x,y) \geq 1$. Cazals *et al.* (2002) showed that the order-m efficiency score $\lambda_m(x,y)$ has an explicit expression that depends only on the conditional distribution $S_Y(y \mid x)$:

$$\lambda_m(x,y) = \int_0^\infty [1 - (1 - S_Y(uy \mid x))^m] du. \tag{8}$$

Similarly with FDH, one can then obtain the estimator for the order-m efficiency by plugging the $\widehat{S}_{Y,n}(y \mid x)$ to equation (8), which gives $\widehat{\lambda}_{m,n}(x,y) = \int_0^\infty [1-(1-\widehat{S}_{Y,n}(uy \mid x))^m] du$. Note that this estimator is relatively easy to compute, as it based on a univariate integral. As shown by Cazals et al. (2002), the remarkable statistical property of the order-m estimator $\widehat{\lambda}_{m,n}(x,y)$ is its $\sqrt{n}$-consistency, i.e. it converges to the true value as quickly as parametric estimators. Since this is valid for the general multiple input-output case, the estimator avoids the curse of dimensionality problem, which is very rare for nonparametric methods.

2.2 Conditional efficiency

Using the probabilistic formulation, Cazals *et al.* (2002) also suggested a conditional efficiency approach which includes external environmental factors that might influence the production process but are neither inputs nor outputs under the control of the producer. Daraio and Simar (2005) extended their ideas to a more general multivariate setup and proposed a practical methodology to evaluate the effect of environmental variables in the production process. As mentioned in the introduction, the main benefit compared to the alternative two-stage

approach is that it can include environmental variables in the efficiency estimates without assuming a separability condition. Indeed, in a favorable operational environment, entities will need less inputs to produce the given set of outputs. Contrarily, an unfavorable operational environment increases the input requirements. Therefore, the exogenous environment definitely influences the input-output selection and its levels. The conditional efficiency approach consists of conditioning the production process to a given value of $Z = z$, where Z denotes variables characterizing the operational environment. The joint probability function given $Z = z$ can be defined as:

$$H_{XY|Z}(x, y \mid z) = \Pr(X \leq x, Y \geq y \mid Z = z). \tag{9}$$

Again, this can be further decomposed into:

$$\begin{aligned} H_{XY|Z}(x, y \mid z) &= \Pr(Y \geq y \mid X \leq x, Z = z)\ \Pr(X \leq x \mid Z = z) \\ &= S_{Y|X,Z}(Y \geq y \mid X \leq x, Z = z)\ F_X(X \leq x \mid Z = z) \\ &= S_Y(y \mid x, z)\ F_X(x \mid z). \qquad \textit{(in shorthand notation)} \end{aligned} \tag{10}$$

The support of $S_Y(y \mid x, z)$ defines the production technology when $Z = z$. To reduce the deterministic nature, again instead of using the full support of $S_Y(y \mid x, z)$ one can draw randomly m variables $Y_i, i = 1, ..., m$ for which $X \leq x$ and use the expected value of these draws as the efficiency measure $\lambda_m(x, y \mid z)$. Analogously to the unconditional order-m efficiencies, $\lambda_m(x, y \mid z)$ can be expressed using the following integral:

$$\lambda_m(x, y \mid z) = \textstyle\int_0^\infty [1 - (1 - S_Y(uy \mid x, z))^m] du. \tag{11}$$

Estimating $S_Y(y \mid x, z)$ nonparametrically is somewhat more difficult than for the unconditional case, as we need to use smoothing techniques in z (due to the equality constraint $Z = z$):

$$\hat{S}_{Y,n}(y \mid x, z) = \frac{\sum_{i=1}^n I(x_i \leq x, y_i \geq y) K_h(z, z_i)}{\sum_{i=1}^n I(x_i \leq x) K_h(z, z_i)}, \tag{12}$$

where $K_h(\cdot)$ is a kernel and h is an appropriate bandwidth for this kernel. The conditional order-m efficiency estimator $\hat{\lambda}_{m,n}(x, y \mid z)$ is then obtained by plugging $\hat{S}_{Y,n}(y \mid x, z)$ into equation (11), i.e.

$$\hat{\lambda}_{m,n}(x, y \mid z) = \textstyle\int_0^\infty [1 - (1 - \hat{S}_{Y,n}(uy \mid x, z))^m] du. \tag{13}$$

Importantly, Cazals *et al.* (2002) showed that the convergence rate of estimator $\widehat{\lambda}_{m,n}(x, y \mid z)$ depends on the dimension of Z, being $(nh^r)^{-1/2}$, where $r = \dim(Z)$.[4] This means that

[4] Here it is assumed that bandwidth is similar for all environmental variables in Z. However, this assumption can be easily relaxed, as we will do later.

although order-m estimator avoids the curse of dimensionality, the accuracy of the conditional estimator depends on the dimension of Z due to the smoothing in z.

The current literature assumes that the univariate/multivariate Z is continuous. Clearly, an extension of the conditional efficiency approach to a more general setting including both discrete and continuous variables requires changes to the presented framework, because in general it is not appropriate to treat discrete variables similarly with continuous (i.e. use continuous kernel for all ordered and unordered discrete variables). Next section discusses the treatment of discrete variables, the choice of kernel functions and the bandwidth selection in a generalized setting including both discrete and continuous exogenous variables.

3 Estimation with mixed data

3.1 Motivation

This section shows how to generalize the conditional efficiency approach to the case of environmental factors having both discrete and continuous components. Firstly, it is important to notice that the conditional efficiency approach presented in Section 2 is similar to traditional nonparametric methods (like kernel methods) used in regression and density estimation with respect to the presumption that the underlying data is continuous. If one would have a data set containing a mix of continuous and discrete data, the conventional approach in nonparametric estimation would be to split the sample in subgroups (or 'cells') corresponding to the different values of the discrete variables and then estimate separate models/functions for those subsamples. This approach is sometimes referred to as a 'frequency-based' method (see e.g. Li and Racine, 2007). One could follow the frequency-based approach also in the conditional efficiency estimation by splitting the sample to subgroups with respect to the values of discrete variables, and then employ the methods presented in Section 2 for each of the subgroups (using inputs, outputs and continuous environmental variables). In essence, this would combine the conditional efficiency approach with the frontier separation approach referred in the introduction.

However, there are some important reasons why we prefer the alternative approach (presented below), which does not require the sample splitting *a priori*. The first and perhaps also the most important reason is that the frequency-based method will be problematic and even infeasible when the sample size is not large relative to the number of subgroups. For example, in our empirical application the sample size is 293, and the number of subgroups (or cells) is $6 \times 6 \times 3 \times 2 \times 16 = 3456$ meaning that there are only $293/3456 \approx 0.08$ observations per subgroup on average! We note that this is not just a curious example; in fact, efficiency applications using parametric regression methods use frequently many discrete variables in relative small samples (100-300 observations). Clearly, one can then not use the nonpara-

metric frequency-based method without ignoring some discrete variables from the analysis. Besides the infeasibility problem, it is not practical to estimate a large number of models for different values of discrete variables. The second relevant disadvantage of the frequency-based method concerns statistical inference. Although it is is quite straigthforward to test the effect of a dummy variable using boostrapping methods by comparing efficiency distributions of separate groups (as Daraio and Simar, 2007a, also mention), the test is much more challenging if there are more than two subgroups and in particular if one wants to test significance of the categorical variable that has many classes.

To avoid the problems of the frequency-based method, we propose to use an alternative approach that smooths also the discrete variables in a particular manner (as first suggested by Aitchison and Aitken, 1976). The idea of smoothing discrete along with continuous variables is based on novel kernel methods presented by Qi Li, Jeff Racine and their colleagues (see e.g. Li and Racine, 2003; Racine and Li, 2004; Hall, Li and Racine, 2004; Li and Racine 2007, 2008). We introduce and adapt these techniques in next subsections to our framework.

3.2 Generalized kernel estimation

As we treat continuous, discrete ordered (i.e. the discrete variables have a meaningful order) and discrete unordered variables (i.e. it does not matter how the variables are classified to categories) differently in the estimations, we redefine the multivariate Z. Define a vector of observed environmental variables by $z_i = (z_i^c, z_i^o, z_i^u)$, $i = 1, ..., n$, where the first component $z_i^c \in \mathbb{R}^r$ denotes a vector of continuous environmental variables, z_i^o is a v-dimensional vector of environmental variables that assume ordered discrete values and z_i^u is a w-dimensional vector of exogeneous variables that assume unordered discrete values. In addition, let z_{is}^o and z_{is}^u denote sth components of z_i^o and z_i^u. Without losing any generality, we assume that z_{is}^o and z_{is}^u can take $c_s \geq 2$ and $d_s \geq 2$ different values, i.e. $z_{is}^o = \{0, 1, ..., c_s - 1\}$ for $s = 1, ..., v$ and $z_{is}^u = \{0, 1, ..., d_s - 1\}$ for $s = 1, ..., w$. This means that the support of z_i^o and z_i^u are $S^o = \prod_{s=1}^{v} \{0, 1, ..., c_s - 1\}$ and $S^u = \prod_{s=1}^{w} \{0, 1, ..., d_s - 1\}$, respectively.

To smooth both continuous and discrete variables, we need to use kernel functions for all the environmental variables. We follow Li and Racine (2007) and use a standard multivariate product kernel for all three components in z_i.[5] By multiplying these multivariate kernel functions, we obtain a generalized product kernel function, formally expressed as:

$$K_h(z, z_i) = \prod_{s=1}^{r} \frac{1}{h_s^c} l^c \left(\frac{z_s^c - z_{is}^c}{h_s^c} \right) \prod_{s=r+1}^{r+v} l^o \left(z_s^o, z_{is}^o, h_s^o \right) \prod_{s=r+v+1}^{r+v+w} l^u \left(z_s^u, z_{is}^u, h_s^u \right), \qquad (14)$$

where $l^c(\cdot)$, $l^o(\cdot)$ and $l^u(\cdot)$ are univariate kernel functions and h_s^c, h_s^o and h_s^u are bandwidths for, respectively, continuous, ordered and unordered environmental variables. Regarding the

[5] Of course, if any of the components z_i^c, z_i^o or z_i^u are univariate, then an univariate kernel suffices for that component.

continuous kernel function $l^c(\cdot)$, we know from the previous research (Daraio and Simar, 2005) that one should use kernels with compact support (i.e. kernels for which $k(z) = 0$ if $|z| \geq 1$) such as the uniform, triangle, Epanechnikov or quartic kernels. In this study we will use the Epanechnikov kernel (although other compact kernels deliver very similar results). For unordered variables we employ the Aitchison and Aitken (1976) discrete univariate kernel function that was designed for discrete variables without any order, while for ordered discrete variables we employ the Li and Racine (2007) discrete kernel function that also takes into account the ordering of the categories. Formally, these continuous and discrete kernel functions are given by:

$$l^c\left(\frac{z_s^c - z_{is}^c}{h_s^c}\right) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}\left(\frac{z_s^c - z_{is}^c}{h_s^c}\right)^2\right) & \text{if } \left(\frac{z_s^c - z_{is}^c}{h_s^c}\right)^2 \leq 5 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

$$l^u\left(z_s^u, z_{is}^u, h_s^u\right) = \begin{cases} 1 - h_s^u & \text{if } z_{is}^u = z_s^u \\ h_s^u/\left(c_s - 1\right) & \text{if } z_{is}^u \neq z_s^u \end{cases} \tag{16}$$

$$l^o\left(z_s^o, z_{is}^o, h_s^o\right) = \left(h_s^o\right)^{|z_{is}^o - z_s^o|}. \tag{17}$$

It is worth considering the two discrete kernel functions in more detail, as they have not been previously used in nonparametric efficiency literature. Firstly, both the Aitchison and Aitken (1976) and Li and Racine (2007) kernel functions impose contraints for bandwidth parameters. For the former, bandwidth h_s^u must be between 0 and $\left(c_s - 1\right)/c_s$, whereas for the latter bandwidth h_s^o can take values between [0,1].[6] By considering the limit values of h_s^u, we see that when $h_s^u = 0$ then $l^u\left(z_s^u, z_{is}^u, 0\right) = I(z_{is}^u = z_s^u)$ becomes an indicator function, while $h_s^u = \left(c_s - 1\right)/c_s$ gives $l^u\left(z_s^u, z_{is}^u, \left(c_s - 1\right)/c_s\right) = 1/c_s$, i.e. a constant kernel function. The first special case is of particular interest, because the indicator function divides the sample to subgroups exactly the same way as the frequency-based method discussed in Section 3.1. Similarly, we can observe that when $h_s^o = 1$, Li and Racine kernel function becomes $l^o\left(z_s^o, z_{is}^o, h_s^o\right) = 1$ for all values of z_s^o and $z_{is}^o \in \{0, 1, ..., c_s - 1\}$ such that the irrelevant variable z_s^o will be smoothed out. In our conditional efficiency setting, the discrete kernel estimations boil intuitively down to in the order-m estimation drawing with a positive probability of $(1 - h_s^u)$ observations which belong to the same class as the evaluated observation, and with a positive probability of $h_s^u/\left(c_s - 1\right)$ (or alternatively for unordered variables $\left(h_s^o\right)^{|z_{is}^o - z_s^o|}$) observations which do not belong to this class. Drawing observations which both belong to and not belong to the evaluated class (although with a different probability) smooths the discrete variable.

Having presented the idea of smoothing the mixed variables with the generalized kernel approach, we apply the technique to the conditional efficiency framework. For multivariate

[6] For example, if we have a unordered dummy variable, we know that $c_s = 2$ and thus $h_s^u \in [0, 1/2]$.

$z = (z^c, z^o, z^u)$ inluding continuous and unordered and ordered discrete components, the estimator for the conditional survivor function of Y can be expressed as:

$$\widehat{S}_{Y,n}(y \mid x, z) = \frac{\sum_{i=1}^{n} I(x_i \leq x, y_i \geq y) K_h(z, z_i)}{\sum_{i=1}^{n} I(x_i \leq x) K_h(z, z_i)}, \tag{18}$$

where $K_h(z, z_i)$ is the generalized multivariate kernel function specified in equation (14). Further, one can again obtain the conditional efficiency estimator $\widehat{\lambda}_{m,n}(x, y \mid z)$ by plugging in $\widehat{S}_{Y,n}(y \mid x, z)$ in equation (8).

To show the validity of the approach, and in particular to show the consistency of the estimators, we make the following assumptions.

Assumption (A1): The sample observations $S_n = \{(x_i, y_i, z_i) \mid i = 1, ..., n\}$ are realizations of independent and identically distributed (iid) random variables (X, Y, Z) with the probability density function $f_{XYZ}(x, y, z)$. Both the marginal density function $f_Z(z)$ and the conditional survivor function $S_Y(y \mid x, z)$ have continuous second order partial derivatives with respect to z^c. For fixed values of x, y and z, $f_Z(z) > 0$ and $0 < S_Y(y \mid x, z) < 1$.

Assumption (A2): $l^c(\cdot)$ is a symmetric, bounded, and compactly supported density function.

Assumption (A3): As $n \to \infty$, $h_s^c \to 0$ for $s = 1, ..., r$, $h_s^o \to 0$ for $s = 1, ..., v$, $h_s^u \to 0$ for $s = 1, ..., w$, and $(nh_1^c h_2^c ... h_r^c)^{-\frac{1}{2}} \to \infty$.

The following theorem and corollary give the convergence rate of $\widehat{S}_{Y,n}(y \mid x, z)$ and $\widehat{\lambda}_{m,n}(x, y \mid z)$.

Theorem 1 *Under Assumptions (A1) to (A3),* $\widehat{S}_{Y,n}(y \mid x, z)$ *converges to* $S_Y(y \mid x, z)$ *with* $O_p\left((nh_1^c h_2^c ... h_r^c)^{-\frac{1}{2}}\right)$.

Proof.

First, note that we can write the conditional survivor function estimator as:

$$\widehat{S}_{Y,n}(y \mid x, z) = \frac{\sum_{i \in N_x} I(y_i \geq y) K_h(z, z_i)}{\sum_{i \in N_x} K_h(z, z_i)}, \tag{19}$$

where $N^x = \{x_i \mid I(x_i \leq x) = 1, i = 1, ..., n\}$. Li and Racine (2008) prove that $\widehat{F}_{Y,n}(y \mid z) = \frac{\sum_{i=1}^{n} I(y_i \leq y) K_h(z, z_i)}{\sum_{i=1}^{n} K_h(z, z_i)}$ converges to $F_Y(y \mid z)$ in mean square error (and hence in probability) with $O_p\left((nh_1^c h_2^c ... h_r^c)^{-\frac{1}{2}}\right)$ under regularity conditions that are similar to Assumptions (A1)-(A3). Besides $X \leq x$, the only difference to Li and Racine (2008) is that we are estimating the conditional survivor function $S_Y(y \mid z)$ instead of the conditional distribution function $F_Y(y \mid z)$. Since by definition $S_Y(y \mid z) = 1 - F_Y(y \mid z)$, their results extends to our case when condition on $X \leq x$. ■

The following result follows directly from Theorem 1, as for given m $\lambda_m(x, y \mid z)$ depends only on $S_Y(y \mid x, z)$.

Corollary 1 *Under Assumptions (A1) to (A3),* $\widehat{\lambda}_{m,n}(x, y \mid z)$ *converges to* $\lambda_m(x, y \mid z)$ *with* $O_p\left((nh_1^c h_2^c ... h_r^c)^{-\frac{1}{2}}\right)$ *for any fixed value of m.*

These results prove that the conditional efficiency estimator $\widehat{\lambda}_{m,n}(x, y \mid z)$ is consistent in a more general case including both discrete and continuous environmental variables. Additionally, they show that the convergence rate of the estimator is $(nh_1^c h_2^c ... h_r^c)^{-\frac{1}{2}}$, i.e. it does not depend on the number of discrete variables in Z but only on the number of continuous variables. This is very relevant result, since efficiency applications use frequently several discrete exogenous factors in small samples.

3.3 Bandwidth selection: A data-driven method

The bandwidth selection is the most crucial step in a nonparametric kernel estimation (cfr. it has the same importance as the model specification in parametric estimations). If the bandwidth is too large, the kernel function will be oversmoothed; if the bandwidth is too small, the kernel function will be undersmoothed. The initial proposal of Daraio and Simar (2005) estimated for z^c the bandwidth h_s^c by the likelihood cross-validation k-nearest neighbor technique. However, (1) only asymptotic optimality of this approach has been shown and (2) although the conditional efficiency estimates try to avoid the separability condition, its bandwidth selection relies on it. Indeed, by only relying on the exogenous variables, the estimation of h_s^c ignores the impact of z^c on the production process (i.e. the impact of z^c on y given that $x_i \leq x$). Therefore, conditional bandwidth estimations are required.

Similar as before, the main challenge lies in extending the traditional bandwidth estimations for y conditional on $Z = z$, to estimations for y conditional on $X \leq x$ and $Z = z$ (as required by the conditional efficiency model). The former conditional bandwidth estimations are developed by the models of Li and Racine (2007, 2008) and Hall *et al.* (2004). The latter conditional efficiency estimations are explored by Badin *et al.* (2008) for continuous variables only. Following the lines of Badin *et al.* (2008) we adopt the model of Hall *et al.* (2004) to a generalized conditional efficiency framework.

Before going more into detail on the approach, we highlight that several procedures for conditional bandwidth estimation exist. For example, the *seemingly* easier plug-in method. It only *seems* easier as plug-in methods could be extremely computational intensive and, more importantly, they do not necessarily lead to an optimal bandwidth if some of the covariates are irrelevant (Li and Racine, 2008). Therefore, we opt for a data-driven approach. Although there does not exist a data-driven bandwidth selection approach for mixed conditional CDF, Li and Racine (2008) suggest to estimate the bandwidth by the least squares cross-validation method based on the closely related conditional probability density functions (PDF), as outlined by Hall *et al.* (2004). As major advantage, the latter procedure removes irrelevant covariates by oversmoothing these variables.

To estimate bandwidths (h^c, h^o, h^u), we minimize the cross-validation function $CV(h^y, h^c, h^o, h^u)$, where h^y is a bandwidth vector for outputs y. Note that although we estimate bandwidths also for y, those bandwidths are not used in conditional efficiency estimation. [7] Define therefore the conditional PDF of Y for $X \leq x$ and $Z = z$ (with $z = (z^c, z^o, z^u)$) as $g(y \mid X \leq x, Z = z) = f(y, X \leq x, Z = z)/m(X \leq x, Z = z)$ where f denotes the joint density of (y, z) and m the marginal density of z for given $X \leq x$. The density f and the marginal density m are not observed but can be estimated by the use of nonnegative, generalized kernels $K(\cdot)$ and $L(\cdot)$:

$$\begin{aligned} \hat{f}(y, x_i \leq x, z) &= \tfrac{1}{n} \textstyle\sum_{i=1}^{n} I(x_i \leq x) K_h(z, z_i) L_{h_y}(y, y_i) \\ \hat{m}(x_i \leq x, z) &= \tfrac{1}{n} \textstyle\sum_{i=1}^{n} I(x_i \leq x) K_h(z, z_i) \end{aligned} \tag{20}$$

where the generalized kernel $K_h(z, z_i)$ is computed as in equation (14) and the multivariate kernel $L_{h_y}(y, y_i)$ as $\prod_{j=1}^{q} \frac{1}{h_{y_j}} l\left(\frac{y_j - y_{ij}}{h_{y_j}}\right)$ with $l(\cdot)$ a univariate kernel function (Epanechnikov).

As also remarked by Badin *et al.* (2008, p. 8), the only difference between the generalized conditional bandwidth computation of Hall *et al.* (2004) and the optimal data-driven bandwidth needed for the conditional efficiency framework is the reduction of the reference sample size where (h^c, h^o, h^u) are computed in. In particular, instead of using the full reference sample (consisting of n observations) we only consider the observations for which $x_i \leq x$ and compute for this limited reference set the bandwidths (h^c, h^o, h^u). As such, we obtain for every observation a particular set of bandwidths in each of its dimensions (i.e. for every element of z_i). As a disadvantage, this approach dramatically limits the number of reference units for observations with a small x.[8]

Following Hall *et al.* (2004), we start from the weighted integrated squared error (ISE) between $\hat{g}(\cdot)$ and $g(\cdot)$:

$$\begin{aligned} ISE \quad &= \textstyle\int \left\{\hat{g}(y \mid x_i \leq x, z) - g(y \mid x_i \leq x, z)\right\}^2 m(x_i \leq x, z) dW(z) dy \\ &= \textstyle\int \hat{g}(y \mid X \leq x, z)^2 m(x_i \leq x, z) dW(z) dy && (I_{1n}) \\ &\quad -2 \textstyle\int \hat{g}(y \mid X \leq x, z) g(y \mid X \leq x, z) m(x_i \leq x, z) dW(z) dy && (I_{2n}) \\ &\quad + \textstyle\int g(y \mid X \leq x, z)^2 m(x_i \leq x, z) dW(z) dy && (I_{3n}) \end{aligned} \tag{21}$$

where $dW(z)$ denotes an infinitesimal element of a measure (in order to avoid for the continuous components of z, z^c, dividing by 0 in the ratio $\hat{f}(y, x_i \leq x, z)/\hat{m}(x_i \leq x, z)$). The leading term of the ISE (i.e. the part depending on the bandwidth; which corresponds in equation (21) with the terms I_{1n} and I_{2n} as these have estimates of $g(\cdot)$) can be approximated by a cross-validation (CV) objective function which does not assume numerical integration, nor

[7] In total, there are $q+r+v+w$ bandwidths: $(h^y, h^c, h^o, h^u) = (h^y_1, ..., h^y_q, h^c_1, ..., h^c_r, h^o_1, ..., h^o_v, h^u_1, ..., h^u_w)$, but only bandwidth vectors h^c, h^o and $h^u$are used in conditional efficiency estimation.

[8] Note that this is also the case for the traditional and robust FDH estimator of, respectively, Deprins *et al.* (1984) and Cazals *et al.* (2002).

initial assumptions on bandwidths or density function estimators. Hall *et al.* (2004) (and extended by Badin *et al.*, 2008) show that the leading term of the CV criterion corresponds to:

$$CV(h_1^y, ..., h_q^y, h_1^c, ..., h_r^c, h_1^o, ..., h_v^o, h_1^u, ..., h_w^u) = \hat{I}_{1n} - 2\hat{I}_{2n} \tag{22}$$

where the empirical approximations of I_{1n} and I_{2n}, respectively, $\hat{I}_{1n}$ and $\hat{I}_{2n}$, are based on a leave-one-out sample, i.e. a sample of $(n-1)$ observations due to deleting observation i from the sample. By optimizing $(h_1^y, ..., h_q^y, h_1^c, ..., h_r^c, h_1^o, ..., h_v^o, h_1^u, ..., h_w^u)$, we minimize the CV function.

It can be shown that the optimal order of the bandwidths corresponds $h_s^c \sim n^{-1/(5+r)}$ and $h_s^{o,u} \sim n^{-2/(5+r)}$ (Li and Racine, 2008). However, as we basically estimate the optimal bandwidth for the conditional PDF instead of for the closely related conditional CDF, we need to adjust the bandwidths to obtain bandwidths of the optimal order of $h_s^c \sim n^{-1/(4+r)}$ and $h_s^{o,u} \sim n^{-2/(4+r)}$. The optimal bandwidths (as computed along the conditional PDF) can be corrected by multiplying h_s^c with $n^{\frac{1}{5+r}-\frac{1}{4+r}}$ and $h_s^{o,u}$ by $n^{\frac{2}{5+r}-\frac{2}{4+r}}$.

Finally, we note that in some applications one might want to compare performance of units only with the observations in the same category (i.e. the same value of discrete variable). For example, in evaluating efficiency of hospitals using data from several countries, one may want to limit comparison units to hospitals in the same country because of the technological and operational differences. In our framework this is very easy to implement by imposing bandwidth to be zero for the discrete variable in question (i.e. country). It is worth emphasizing that the presented framework still allow bandwidths of other discrete environmental variables to be positive and is therefore more general than the nonparametric frontier separation (or metafrontier) approach.

3.4 Examining the influence of exogenous variables on the production process

3.4.1 Visualization

To evaluate systematically the influence of exogeneous variables on the production process, we compare the conditional efficiency measure $\widehat{\lambda}_{m,n}(x, y \mid z)$ with the unconditional efficiency measure $\widehat{\lambda}_{m,n}(x, y)$. In particular, we follow the methodology suggested by Daraio and Simar (2005, 2007a) by nonparametrically regressing the ratio of the conditional and unconditional efficiency measure $Q^z = \frac{\widehat{\lambda}_{m,n}(x,y|z)}{\widehat{\lambda}_{m,n}(x,y)}$ to environmental factors z. They use a smooth nonparametric kernel regression to estimate the model $Q_i^z = f(z_i) + \epsilon_i$. In addition, they visualize the estimated relationships between environmental variables and the ratio of efficiency scores. Using simulations, Daraio and Simar showed that this approach allows one to detect positive,

negative, neutral or even nonmonotone effects of the environmental factors on the production process.

When Z is continuous and univariate the visualization is straigthforward as one can use scatterplots of Q^z against Z, and as a smoothed nonparametric regression curve can illustrate the effect of Z on Q^z. For example in an output-oriented efficiency, a horizontal line implies that Z does not affect the production process, whereas an increasing (decreasing) smoothed regression curve shows that Z is favorable (unfavorable) to the production process. By interpretation, a favorable effect means that the environmental variable plays the role of a 'substitutive' input in the production process by increasing the productivity of traditional inputs, whereas an unfavorable effect implies that the environmental variable contraints the production by using more inputs in production activity.

When Z is multivariate and includes also discrete variables, visualization is also feasible, although somewhat more challenging. For $\dim(Z) = 2$, one can use 3-dimensional plots. However, if $\dim(Z) > 2$, those are not enough. Perhaps the easiest solution for multivariate cases is to examine so-called *partial regression plots* (see e.g. Daraio and Simar, 2007a; Badin *et al.*, 2008), where only one (or two) environmental variable(s) is (are) allowed to change and other variables are kept at a fixed value. Further, one can then use several different fixed values such as median and 1st and 3rd quartile to examine whether the effect on individual variable Z_s is the same for different values of others exogenous factors. This kind of procedure helps to recognize the effect of individual variable on the production process and possible interactional effects between environmental variables. Moreover, it can be used also for discrete variables as we illustrate in the empirical application.

3.4.2 Nonparametric estimation and inference

Although it can be useful to visualize the effect of environmental variables to the production process, researchers are usually more interested in their statistical significance. Yet in the conditional efficiency framework, so far, only descriptive analysis has been suggested and applied in studying the effect of environmental variables on the production process. This is in sharp contrast to the papers using two-stage models, where tools of statistical inference have been used extensively. Our aim is to propose for robust (thus order-m) conditional efficiency models a framework to test the significance of mixed multivariate environmental variables in the production process. We follow the lines of earlier research by focussing on smoothed nonparametric regression. However, instead of Nadaraya-Watson kernel regression, which has been mostly used in previous conditional efficiency studies, we will use local linear regression for estimating $Q_i^z = f(z_i) + \epsilon_i$. Compared to the Nadaraya-Watson kernel estimator (i.e. local constant regression), the local linear estimator is less sensitive to boundary effects and can also simultaneously uncover the marginal effects of the environmental variables on

Q^z.[9]

As in our framework Z can include both discrete and continuous variables, it is again useful to employ smoothing techniques which allow one to estimate the nonparametric regression model without sample splitting (i.e. which was the case in the frequency approach). Therefore, we use the nonparametric regression method developed by Racine and Li (2004) and Li and Racine (2004), which smooths both continuous and discrete variables. To present the basic idea shortly, consider our nonparametric model:

$$Q_i^z = f(z_i) + \epsilon_i, \qquad i = 1, ..., n \tag{23}$$

where as previously $Q_i^z = \frac{\widehat{\lambda}_{m,n}(x_i, y_i|z_i)}{\widehat{\lambda}_{m,n}(x_i, y_i)}$, $z_i = (z_i^c, z_i^o, z_i^u)$ includes values of continuous, ordered and unordered exogenous variables for observation i, ϵ_i is the usual error term with $E(\epsilon_i | z_i) = 0$, and f is the conditional mean function. The local linear method is based on the following minimization problem:

$$\min_{\{\alpha,\beta\}} \sum_{i=1}^{n} (Q_i^z - \alpha - (z_i^c - z^c)\beta)^2 K_h(z, z_i), \tag{24}$$

where α and β are local coefficients and K_h is the generalized product kernel function defined earlier. Letting $\widehat{\alpha} = \widehat{\alpha}(z)$ and $\widehat{\beta} = \widehat{\beta}(z^c)$ denote the solutions that minimize equation (24), it is straigthforward to show that local linear estimators $\widehat{\alpha}(z)$ and $\widehat{\beta}(z^c)$ are consistent estimators for $f(z) = E(Q^z | z)$ and $\beta(z^c)$. Note that the practical advantage of local linear regression is the fact that one can estimate simultaneously both the conditional mean function $f(z)$ and the gradient vector $\beta(z^c)$ for continuous components (which can be interpreted as varying coefficient). For bandwidth choice we use again the least-squares cross-validation, although one can employ also other methods available in literature. For more details on nonparametric regression with mixed data and the bandwidth choice methods, see Racine and Li (2004) and Li and Racine (2004, 2007).

Before presenting the statistical inference tools, it is important to justify our approach. We want to emphasize that our framework does not suffer from similar inference problems as the two-stage models with the traditional and deterministic FDH and DEA models. Simar and Wilson (2007) rigourously show that most studies using two-stage models have used tools that are invalid in that context. Therefore, to make accurate and valid inference, Simar and Wilson (2007) suggested boostrapping methods, which has been lately used in many applications. There are basically four reasons why our approach avoids the problems listed by Simar and Wilson (2007). Firstly, since our dependent variable Q_i^z is based on the ratio of conditional and unconditional efficiency score, it is not restricted to interval $[0, 1]$ or $[1, \infty)$ (which is the case in traditional deterministic FDH and DEA). Secondly, as the ratio of efficiency scores

[9] Jeong *et al.* (2008) use also local linear procedure to estimate the effect of environmental variable(s).

in Q_i^z can be very different for different observations, there is no reason to suspect that there would be a systematic correlation between observations if $Q_i^z \neq 1$. Moreover, even in the very unlikely case when the conditional and unconditional efficiency scores would be the same for all observations, the possible correlation is a smaller problem (and disappears more quickly) in the robust order-m than in the traditional deterministic FDH and DEA. Thirdly, since our framework does not assume separability (in contrast to two-stage models), there is no reason for ϵ and Z to be systematically correlated. Fourthly, although our estimation methods cannot avoid bias in small samples (similarly to other nonparametric methods), if the number of continuous variables in Z is small our fully nonparametric estimation method has much faster convergence rate than typical semiparametric two-stage models. Therefore, we can conclude that the robust (order-m) conditional efficiency framework does not suffer from the problems in the traditional two-stage model.

Since our estimation framework is fully nonparametric, we also want to avoid any parametric assumptions in the statistical inference stage. It is worth emphasizing that parametric assumptions would be difficult to justify in this context and even inconsistent with our nonparametric efficiency estimation. Thus, to test the significance of regressors in (23), we will utilize recently developed nonparametric tests. More specifically, we test the significance of each of the discrete and each of the continuous variables using tests, respectively, proposed by Racine *et al.* (2006) and Racine (1997). These tests can be seen as the nonparametric equivalent of standard t-tests in ordinary least squares regression. However, nonparametric test are more general than standard t-tests, as the former tests both linear and (unspecified) non-linear relationships. In a multivariate setting the null hypotheses for testing continuous and discrete (both ordered and unordered) components are, respectively:

$$H_0 \quad : \quad E\left(Q^z \left| \widetilde{Z}, Z_s^c \right.\right) = E\left(Q^z \left| \widetilde{Z} \right.\right) \text{ almost everywhere, and} \tag{25}$$

$$H_0 \quad : \quad E\left(Q^z \left| \widetilde{Z}, Z_s^d \right.\right) = E\left(Q^z \left| \widetilde{Z} \right.\right) \text{ almost everywhere,} \tag{26}$$

where Z_s^c and Z_s^d denote sth component of continuous and discrete (ordered or unordered) variables and $\widetilde{Z}$ represent all other environmental variables, which can be both continuous and discrete. The alternative hypotheses H_1 are negations for the null hypotheses. Thus, e.g., for the second case the alternative hypothesis is $H_1 : E\left(Q^z \left| \widetilde{Z}, Z_s^d \right.\right) \neq E\left(Q^z \left| \widetilde{Z} \right.\right)$.

To deduce a practical implementation, we firstly rewrite the null hypothesis for continuous variables as:

$$H_0 : \frac{\partial E\left(Q^z \left| \widetilde{Z}, Z_s^c \right.\right)}{\partial Z_s^c} = \beta\left(Z_s^c\right) = 0 \text{ almost everywhere,} \tag{27}$$

i.e., that the partial derivative of $f(Z)$ with respect to Z_s^c is zero. Using this representation, the test statistic for continuous components can be written as:

$$I^c = E\left\{\beta\left(Z_s^c\right)^2\right\}. \tag{28}$$

A consistent estimator for this test statistic can be obtained by substituting the local linear estimator for unknown derivative and using a sample average of I, i.e.

$$I_n^c = \frac{1}{n}\sum_{i=1}^{n} \widehat{\beta}\left(z_{is}\right)^2. \tag{29}$$

To estimate the finite-sample distribution and critical value of the test statistic I_n^c, nonparametric bootstrap procedures can be used. We shortly explain the steps of the bootstrap procedure; for more details, see Racine (1997). First estimate the conditional mean function $E\left(Q^z \left| \widetilde{Z}, Z_s^c \right.\right) \equiv f^0$ and save residuals $\widehat{\epsilon}_i$, $i = 1, ..., n$. Secondly, resample with replacement from the residual distribution $\widehat{F}$, which has probability mass $\frac{1}{n}$ for all $\widehat{\epsilon}_i$, to obtain a bootstrap sample $\{\widehat{\epsilon}_i^*\}_{i=1}^n$. Thirdly, generate a bootstrap sample $\left\{\widehat{Q}_i^*, z_i\right\}_{i=1}^n$, where $\widehat{Q}_i^* = \hat{f}_i^0 + \widehat{\epsilon}_i^*$, $i = 1, ..., n$ and z_i include all conditioning variables. Fourthly, estimate $\widehat{\beta}\left(z_{is}\right)^*$ and the test statistic using the bootstrap sample. By repeating steps (1)-(4) B times (where B is a large number) one obtains a sample distribution that can be then used for calculating critical values and p-values for the test statistic.

Secondly, for discrete variables a statistic similar to (29) can be used for the significance testing. Let us assume that the testable discrete variable Z_s^d (ordered or unordered) takes c different values, $\{0, 1, 2, ..., c-1\}$. If we denote the conditional mean function by $f(\widetilde{Z}, Z_s^d)$, the null hypothesis $E\left(Q^z \left| \widetilde{Z}, Z_s^d \right.\right) = E\left(Q^z \left| \widetilde{Z} \right.\right)$ is equivalent to $f(\widetilde{Z}, Z_s^d = l) = f(\widetilde{Z}, Z_s^d = 0)$ for all $\widetilde{Z}$ and for $l = 1, 2, ..., c-1$. The test statistic is:

$$I^d = \sum_{l=1}^{c-1} E\left\{\left[f(\widetilde{Z}, Z_s^d = l) - f(\widetilde{Z}, Z_s^d = 0)\right]^2\right\}, \tag{30}$$

which is clearly always nonnegative and equals zero when the null hypothesis is true. A consistent estimator of the test statistic is then obtained as:

$$I_n^d = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{c-1}\left[\widehat{f}(\widetilde{z}_i, z_{is}^d = l) - \widehat{f}(\widetilde{z}_i, z_{is}^d = 0)\right]^2, \tag{31}$$

where $\widehat{f}$ is the local linear estimator of the conditional mean function at the given values of the variables. This estimator can be straightforwardly generalized also to the case, where multiple discrete variables are tested simultaneously (see Racine *et al.*, 2006).

To approximate the finite-sample distribution of I_n^d, Racine *et al.* (2006) suggest to use a boostrap procedure.[10] As the procedure is a bit different than for continuous variables, we next sketch shortly the steps. Firstly, randomly select z_{is}^* from $\{z_{is}\}_{i=1}^n$ with replacement and call $\left\{\widehat{Q}_i^*, \widetilde{z}_i, z_{is}^*\right\}_{i=1}^n$ the bootstrap sample. Secondly, use the bootstrap sample to compute the bootstrap statistic $I_n^{*,d}$, which is otherwise similar than (31) but z_{is} is replaced by z_{is}^*.

[10] Note that Racine *et al.* (2006) propose for discrete variables also two alternative bootstrap procedures that could be used in this context. However, the computational burden is larger.

Thirdly, by repeating steps 1 and 2 B times (with B a large number) one obtains a sample distribution that can be then used for calculating critical values and p-values.

4 Educational efficiency

4.1 The performance of pupils

Our conditional efficiency model allows one to proxy the exogenous environment by a combination of discrete, both ordered and unordered, and continuous variables. The use of combined discrete and continuous variables is particularly valuable when assessing educational data.[11]

We estimate the performance of British pupils at the age of 15 as surveyed by the international Pisa (Program for International Student Assessment) data set for 2006. The latter OECD survey is currently at its third wave (2000, 2003 and 2006) and contains survey data for more than 400,000 pupils from 57 countries. Besides a pupil survey, it consists of a survey by the school and by the parents which try to capture the socio-economic background of the pupil. We limited our sample to 16 randomly chosen English and Welsh schools which count in total 293 surveyed pupils. By considering a small sample, we try to illustrate that our conditional efficiency approach is able to include a large number of discrete variables without losing accuracy of the estimation. As the conditional efficiency model relies on the robust efficiency estimates, it is also well suited to deal with the extremal and atypical observations which could arise from survey data (e.g. Bound *et al.*, 2001).

The conditional order-m estimation requires the selection of input, output and environmental variables. We follow the education literature in selecting these. Students are spending resources (in particular time) to study languages, math, science and other skills. The four input variables sum for, respectively, language, math, science and other subjects the total hours that pupil reported to spend on the subject during regular classes, out of school and self study (i.e. the sum of the variables ST31Q in the Pisa data set). As such, the inputs proxy the devotion to the subjects. Given these efforts, students are obtaining test results which are proxied by 5 plausible values for, respectively, language, math and science (the plausible values are standardized across the OECD countries with an average score of 500). Following the standard literature (e.g. OECD, 2007) we consider as output variables the arithmetic average of the 5 plausible values in the Pisa data set for each of the three subjects. The socio-economic environment (SEE) of the pupil is captured by 7 environmental variables (following Hampden-Thompson and Johnston, 2006 and references therein). We include two ordered variables, i.e. the education of the mother and the father as proxied by

[11] Obviously, the scope of the generalized conditional efficiency framework is much broader. Therefore, the *R* code is available from the authors upon request. The code utilizes *np* package by Hayfield and Racine (2008).

Table 1: Descriptive statistics

		Minimum	Median	Mean	Maximum	St. Dev.
Input	Hours devoted to language	0	6	6	21	3
	Hours devoted to math	0	6	6	21	3
	Hours devoted to science	0	6	6	13	3
	Hours devoted to other subject	0	7	8	21	4
Output	Test score language	214	477	474	673	90
	Test score math	246	472	474	667	74
	Test score science	227	487	492	715	78
SEE	Education mother	1	4	4	6	1
	Education father	1	4	4	6	1
	Lang. at home (1=diff; 2=other nat; 3=Eng)					
	Own room (1=No; 2=Yes)					
	School					
	School size	187	1003	946	1501	326
	Students per teacher	12	16	15	17	1

a variable between 0 (did not complete ISCED 1; where ISCED denotes the International Standard Classification of Education by the Unesco) and 6 (completed ISCED 5a or 6). We also condition on three unordered variables: whether the language at home is the test language (denoted by a value of 3), another national language (a value of 2) or another language (a value of 1); whether the pupil possesses his/her own room (with a value of 2 if so, 1 if not); and a factor denoting the school. The latter variable captures the clustering at the school level which could, e.g., arise from the neighborhood the school is located. Finally, we include two continuous variables which are related to the school characteristics: the total school size and the average teacher-student ratio of the school. Some descriptive sample statistics are presented in Table 1.

Following Daraio and Simar (2005, 2007a) we select the size of the partial frontier m as the value for which the percentage of super-efficient observations (i.e. $\lambda^m_{FDH} < 1$) remains more or less stable. In sample under study, m corresponds to 30. The R code, using some features of the np-package of Hayfield and Racine (2008), is available upon request.

4.2 Results

To assess the performances of the pupils, we estimate the extent to which the pupils are able to deploy their acquired knowledge to obtain higher test results (i.e. an output-orientation). Using this input and output set, we experimented with various combinations of the exogenous

variables. As in all models the discrete variables had a significant effect on the performance of the pupils, we present only two models and particularly discuss the model with only school size as a continuous variable. Denote 'Model 1' as the general model with all exogenous variables, and 'Model 2' as the model without student-teacher ratio. Applying a standard robust order-m model (so without taking the exogenous environment into account), we obtain average efficiency scores of $\lambda^m_{FDH}(x, y) = 1.22$ (see also Table 2). This indicates that if all pupils would perform as efficient as the best practice pupils (i.e. those pupils who are obtaining with a given devotion to the subjects the highest test results), the test scores could on average increase by 22%. Note that some pupils have an efficiency score below 1. These 'super-efficient' pupils are performing better than the average m ($m = 30$) pupils they were benchmarked with in the order-m procedure. Obviously, these efficiency scores are influenced by the socio-economic background of the pupils. We try to capture the pupil and school specific background by a mix of 7 discrete and continuous exogenous variables (Model 1). Taking into account pupil and school characteristics, the conditional efficiency scores reduce to $\lambda^m_{FDH}(x, y \mid z) = 1.15$. By excluding the number of students per teacher as exogenous variable $\lambda^m_{FDH}(x, y \mid z)$ reduces to 1.14 (Model 2). Summary statistics for the pupil-specific bandwidth estimates in Model 2 are presented in Table 2. We observe that the bandwidth for the school size is very large. This is a result of effectively smoothing out the insignificant variables. On the contrary, the discrete variables have rather narrow bandwidths which seem to indicate their significant influence on the production process. This will be tested next.

To examine the influence (i.e. favorable or unfavorable) of the exogenous variables, we nonparametrically regress the exogenous variables on the ratio of the conditioned to the unconditioned efficiency scores. From examining the significance tests and the partial regression plots for the discrete and continuous variables (see below), we learn that the average effect on efficiency is positive and significantly different from 0 for all discrete variables and insignificantly negative for the continuous variables (see Table 3). The average favorable effect for the first two variables (education of mother and father) means that for median values of the other variables, the effect is positive. This means that the larger z the more the unconditioned efficiency score will benefit from z if it is favorable (and thus the higher the ratio). Instead, for unordered discrete variables we cannot give similar interpretation, as classes do not have natural ordering. However, we can see whether there are significant differences between classes and which classes are favorable for educational efficiency. Overall, our results are in line with the general (parametric) literature (see Sirin (2005) for a comprehensive overview of published articles between 1990 and 2000):

- More educated parents will stimulate and encourage their children, such that for a given study devotion these will obtain higher test results.

Table 2: Efficiency estimates and bandwidth

	Minimum	Median	Mean	Maximum	St. Dev.
Unconditional eff.	0,9316	1,1974	1,2160	2,0270	0,1867
Conditional eff. - Model 1	0,9993	1,1028	1,1466	1,9174	0,1571
Conditional eff. - Model 2	0,9998	1,0905	1,1384	1,8803	0,1518
Bw education mother (M2)	0,0000	0,4514	0,4407	0,6848	0,1265
Bw education father (M2)	0,0001	0,3269	0,3409	0,6848	0,1924
Bw lang. at home (M2)	0,0000	0,1538	0,1573	0,4210	0,1323
Bw own room (M2)	0,0000	0,1770	0,1864	0,3424	0,1185
Bw school effect (M2)	0,0000	0,6075	0,5665	0,6420	0,1364
Bw school size (M2)	8,275E-05	5,042E+09	7,321E+09	9,975E+10	8,457E+09

Table 3: Nonparametric significance test

	Model 1 p-value	Model 2 p-value	Average effect as revealed from partial plot	Interpretation
Education mother	0.075 *	0.079 *	Favorable	Higher education is better
Education father	0.012 **	0.015 **	Favorable	Higher education is better
Language	0.012 **	0.016 **	-	Same language is better
Own room	0.041 **	0.008 ***	-	Own room is better
School variable	0.154	0.032 **	-	Effect between schools
School size	0.153	0.155	Unfavorable	Smaller school is better
Student-teacher ratio	0.510		Unfavorable	Smaller classes are better

where "***" denotes significance at 1% level, "**" at 5% and "*" at 10%.

- Children which are facing language difficulties at school (because they speak a different language at home) obtain for a given effort lower test results.

- Besides creating a good study environment, the possession of an own room can proxy the wealth of the family. Pupils with an own room (or, alternatively, from a wealthier family) obtain better results.

- There are significant differences between schools. This school variable can proxy the neighborhood effects and clustering of pupils (which is in line with the metafrontier literature on school and pupil decompositions, see Thanassoulis and Portela, 2002 and references therein).

As mentioned above we use partial regression plots (see Section 3.4) to visualize the effect of the exogenous environment. In a generalized multivariate framework, we set all other exogenous variables on their median value (and, respectively, on their first and third quartile value to capture the heterogeneity among pupils) while the discrete variables are evaluated

Table 4: Evaluation of general exogenous variables - example for native language

Constant variable			
Education mother	4	4	4
Education father	4	4	4
Own room	2	2	2
School variable	71	71	71
School size	1003	1003	1003
Evaluation			
Language	1	2	3
1 quartile	0.973	0.921	0.979
Mean	0.934	0.937	0.938
3 quartile	0.878	0.910	0.919

once at each category (continuous variables are evaluated at 50 evaluation points). Here we illustrate the approach for the native language and for the school size. While keeping all other exogenous variables at their median value (respectively at their first and third quartile value), we evaluate the variable (*in casu* the language) at its different data points (i.e. factors between 1, representing other language than any national language, and 3 the native language is the same as the test language). The results for the language are presented in Table 4 and in Figure 1 and, respectively, for the school size in Figure 2. Similar as in Daraio and Simar (2005, 2007a), the upward sloping trend points to the favorable effect of the exogenous variables (although this is a multivariate framework where all other observations are hold constant at their median, respectively, first and third quartile value). In contrast to Daraio and Simar (2007) by the use of the nonparametric tests, we are also able to examine the significance level of the exogenous characteristics.

5 Conclusion

This paper introduced mixed continuous and discrete exogenous variables in a conditional efficiency framework. The latter accounts, in estimating relative efficiency scores, for heterogeneity among the evaluated entities without assuming a separability condition (i.e. the environmental variables do not affect the level of the inputs and outputs). We explored the probabilitistic framework where it is relying on. However, the traditional conditional efficiency model faced two main drawbacks. Firstly, it has only been developed for continuous exogenous variables. In more interesting real life applications, the researcher wants to investigate the performance of entities while accounting for a broad set of exogenous variables, including both continuous and discrete variables. By using insights from recent

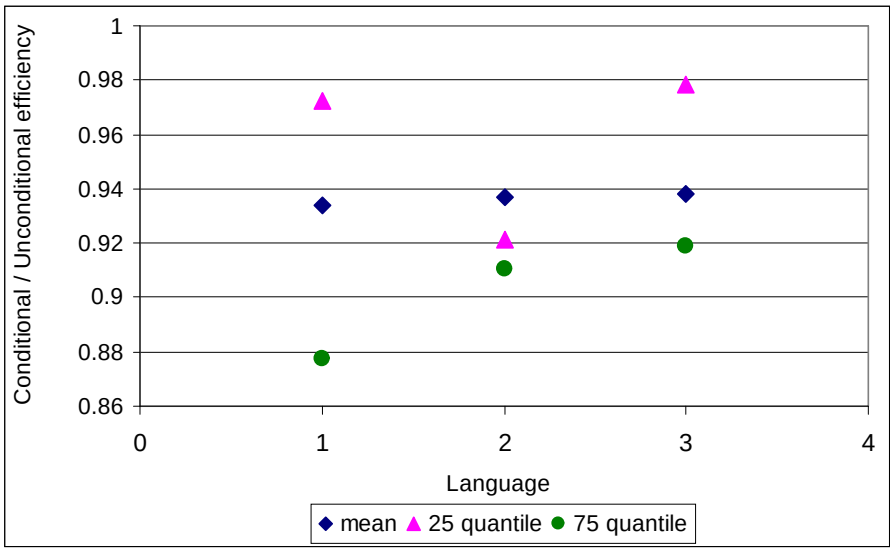

Figure 1: Nonparametric plot of language

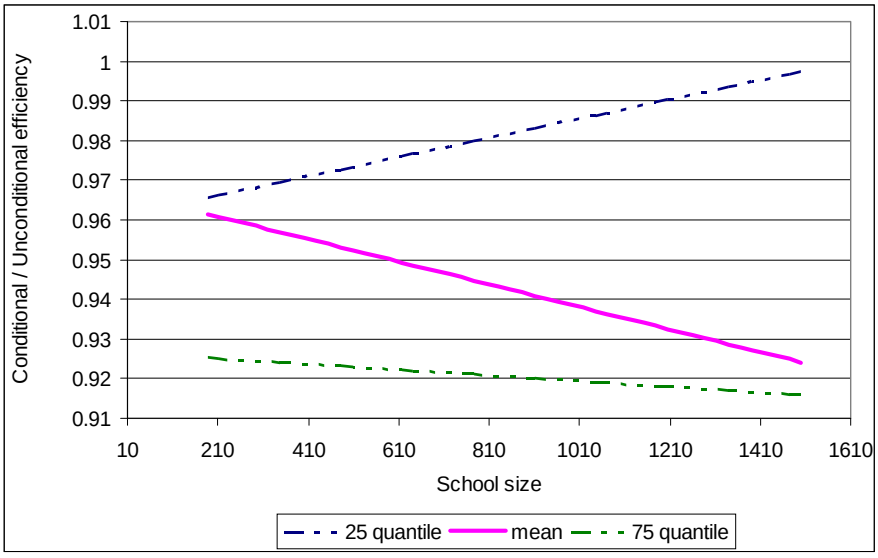

Figure 2: Nonparametric plot of the effect of school size

nonparametric econometrics literature, we generalized the conditional efficiency model to mixed heterogeneous variables. Moreover, we prove that in our setting the discrete component does not suffer from the dimensionality problem, which is the case for continuous environmental variables. Therefore, one can include a large number of discrete environmental variables without reducing the accuracy of the estimation. Secondly, apart from analyzing some descriptive figures, no statistical inference tools have been used in previous studies to test the significance of the exogenous variables. Based on appropriate nonparametric econometric tests, we presented bootstrap procedures for testing the significance of continuous and discrete environmental variables in the production process. In contrast to inference based on more traditional two-stage models, these tests can be used without assuming separability and without any parametric functional forms.

The suggested generalized approach was illustrated on a sample of the OECD Pisa data set. In particular, we examined the performance of British secondary school pupils while taking into account a broad range of continuous, ordered discrete and unordered discrete variables. We find a significant impact on the educational process of each of the discrete exogenous variables. This illustrates that in conditional efficiency estimation one should not limit only to continous environmental variables, but also control for the heterogeneity resulting from the ordered and unordered discrete exogenous factors.

References

[1] Aitchison, J. and C.B.B. Aitken (1976). Multivariate Binary Discrimination by kernel Method. *Biometrika* 63, 413-420.

[2] Badin, L., C. Daraio and L. Simar (2008). Optimal Bandwidth Selection for Conditional Efficiency Measures: A Data-Driven Approach. *Discussion Paper 0828, Institut de Statistique, UCL.*

[3] Banker, R. and R. Morey (1986a). Efficiency Analysis for Exogenously Fixed Inputs and Outputs. *Operations Research* 34, 513–521.

[4] Banker, R. and R. Morey (1986b). The Use of Categorical Variables in Data Envelopment Analysis. *Management Science* 32(12), 1613–1627.

[5] Battese, G., D. Rao and C. O'Donnell (2004). A Metafrontier Production Function for Estimation of Technical Efficiencies and Technology Gaps for Firms Operating under Different Technologies. *Journal of Productivity Analysis* 21 (1), 91-103.

[6] Blass Staub, R. and G. da Silva e Souza (2007). A Probabilistic Approach for Assessing the Significance of Contextual Variables in Nonparametric Frontier Models: an Appli-

cation for Brazilian Banks. *Working Paper Series 150, Central Bank of Brazil, Research Department.*

[7] Bonaccorsi, A., C. Daraio and L. Simar (2007a). Efficiency and Productivity in European Universities. Exploring Trade-Offs in the Strategic Profile. In Bonaccorsi, A. and C. Daraio (eds.). *Universities and Strategic Knowledge Creation. Specialization and Performance in Europe*. Specialization and Performance in Europe. Edward Elgar PRIME Collection.

[8] Bonaccorsi, A., C. Daraio, T. Räty and L. Simar (2007b). Efficiency and University Size: Discipline-Wise Evidence From the European Universities. In: *Hyvinvointipalvelujen Tuottavuus: Tuloksia opintien varrelta*, VATT Publications, Helsinki, Finland.

[9] Bonaccorsi, A. and C. Daraio (2007c). Measuring Knowledge Spillover Effects via Conditional Nonparametric Analysis. Paper presented at the *Workshop on Agglomeration and Growth in Knowledge-based Societies* in Kiel, Germany, April 20-21.

[10] Bonaccorsi A. and C. Daraio (2008). The Differentiation of the Strategic Profile of Higher Education Institutions. New Positioning Indicators Based on Microdata. *Scientometrics* 74 (1), 15-37.

[11] Bonaccorsi, A., C. Daraio and L. Simar (2006). Advanced Indicators of Productivity of Universities, An Application of Robust Nonparametric Methods to Italian Data. *Scientometrics* 66 (2), 389-410.

[12] Bound, J., C. Brown and N. Mathiowetz (2001). Measurement Error in Survey Data. In Heckman J. and E. Leamer (Ed.), *Handbook of Econometrics* 5. Elsevier.

[13] Broekel, T. (2008). From Average to the Frontier: A Nonparametric Frontier Approach to the Analysis of Externalities and Regional Innovation Performance. *Papers in Evolutionary Economic Geography* 08.04.

[14] Broekel, T. and A. Meder (2008). The Bright and Dark Side of Cooperation for Regional Innovation Performance. *Jena Economic Research Papers in Economics 2008-053.*

[15] Cazals, C., J.P. Florens and L. Simar (2002). Nonparametric Frontier Estimation: A Robust Approach. *Journal of Econometrics* 106 (1), 1-25.

[16] Cazals, C, P. Dudley, J.-P. Florens, S. Patel and F. Rodriguez (2008). Delivery Offices Cost Frontier: A Robust Non Parametric Approach with Exogenous Variables. *The Review of Network Economics* 7 (2), 294-308.

[17] Charnes, A., W.W. Cooper and E. Rhodes (1978). Measuring Efficiency of Decision-Making Units. *European Journal of Operational Research* 2 (6), 428-449.

[18] Charnes, A., W.W. Cooper and E. Rhodes (1981). Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to Program follow through. *Management Science* 27 (6), 668-697.

[19] Cherchye, L., K. De Witte, E. Ooghe and I. Nicaise (2007). Equity and Efficiency in Private and Public Education: A Nonparametric Comparison. *CES Discussion Paper Series* DPS 07.25.

[20] Daouia, A. and L. Simar (2007). Nonparametric Efficiency Analysis: A Multivariate Conditional Quantile Approach. *Journal of Econometrics* 140, 375–400.

[21] Daraio, C. and L. Simar (2005). Introducing Environmental Variables in Nonparametric Frontier Models: A Probabilistic Approach. *Journal of Productivity Analysis* 24 (1), 93–121.

[22] Daraio, C. and L. Simar (2006). A Robust Nonparametric Approach to Evaluate and Explain the Performance of Mutual Funds. *European Journal of Operations Research* 175 (1), 516-542.

[23] Daraio, C. and L. Simar (2007a). *Advanced robust and nonparametric methods in efficiency analysis. Methodology and applications*. Series: Studies in Productivity and Efficiency, Springer.

[24] Daraio, C. and L. Simar (2007b). Conditional Nonparametric Frontier Models for Convex and Nonconvex Technologies: A Unifying Approach. *Journal of Productivity Analysis* 28, 13-32.

[25] Deprins, D., L. Simar and H. Tulkens (1984). Measuring Labor Efficiency in Post Offices. In Marchand M., P. Pestieau and H. Tulkens (eds.), *The Performance of Public Enterprises: Concepts and Measurements*. Amsterdam, North-Holland. pp. 243-267.

[26] De Witte, K. and E. Dijkgraaf (2007). Mean and Bold? On Separating Merger Economies from Structural Efficiency Gains in the Drinking Water Sector. *Tinbergen Institute Discussion Paper* 092/3. Accepted for publication in the *Journal of the Operational Research Society*.

[27] De Witte, K. and R. Marques (2008a). Capturing the environment, a metafrontier approach to the drinking water sector. *CES Discussion Paper Series* DPS 08.04; Accepted for publication in *International Transactions of Operational Research*.

[28] De Witte, K. and R. Marques (2008b). Big and beautiful? On non-parametrically measuring scale economies in non-convex technologies. *CES Discussion Paper Series* DPS 08.22.

[29] De Witte, K. and D. Saal (2008). The regulator's fault? On the effects of regulatory changes on profits, productivity and prices in the Dutch drinking water sector. *CES Discussion Paper Series* DPS 08.28.

[30] Farrell, M. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society* 120 (3), 253-290.

[31] Färe, R., S. Grosskopf, C.A.K. Lovell and C. Pasurka (1989). Multilateral Productivity Comparisons When Some Outputs Are Undesirable: A Nonparametric Approach. *The Review of Economics and Statistics* 71(1), 90–98.

[32] Ferrier, G.D. and C.A.K. Lovell (1990). Measuring Cost Efficiency in Banking: Econometric and Linear Programming Evidence. *Journal of Econometrics* 46, 229-245.

[33] Fried, H.O., S.S. Schmidt and S. Yaisawarng (1999). Incorporating the Operating Environment into a Nonparametric Measure of Technical Efficiency. *Journal of Productivity Analysis* 12, 249–267.

[34] Fried, H.O., C.A.K. Lovell, S.S. Schmidt and S. Yaisawarng (2002). Accounting for Environmental Effects and Statistical Noise in Data Envelopment Analysis. *Journal of Productivity Analysis* 17, 157–174.

[35] Fried, H., C.A.K. Lovell and S. Schmidt (2008). *The Measurement of Productive Efficiency and Productivity Growth*. Oxford University Press, pp. 638.

[36] Hall, P., J.S. Racine and Q. Li (2004). Cross-Validation and The Estimation of Conditional Probability Densities. *Journal of the American Statistical Association* 99 (468), 1015–1026.

[37] Hampden-Thompson G. and J. Johnston (2006). Variation in the Relationship between Nonschool Factors and Student Achievement on International Assessments. *National Center for Education Statistics: Statistics in Brief*, NCES 2006-014.

[38] Hayfield, T. and J.S. Racine (2008). Nonparametric Econometrics: The np Package. *Journal of Statistical Software* 27 (5).

[39] Jeong, S., B. Park and L. Simar (2008). Nonparametric Conditional Efficiency Measures: Asymptotic Properties. *Annals of Operations Research*. Forthcoming.

[40] Li, Q. and J.S. Racine (2003). Nonparametric Estimation of Distributions with Categorical and Continuous Data. *Journal of Multivariate Analysis* 86 (2), 266–292.

[41] Li, Q. and J.S. Racine (2004). Cross-Validated Local Linear Nonparametric Regression. *Statistica Sinica* 14(2), 485–512.

[42] Li, Q., and J.S. Racine (2007). *Nonparametric Econometrics: Theory and Practice.* Princeton University Press.

[43] Li, Q. and J.S. Racine (2008). Nonparametric Estimation of Conditional CDF and Quantile Functions with Mixed Categorical and Continuous Data. *Journal of Business and Economic Statistics* 26 (4), 423-434.

[44] Park, B., L. Simar and V. Zelenyuk (2008). Local Likelihood Estimation of Truncated Regression and its Partial Derivatives: Theory and Application. *Journal of Econometrics* 146 (1), 185–198.

[45] Racine, J.S. (1997). Consistent signicance testing for nonparametric regression. *Journal of Business and Economic Statistics* 15 (3), 369-379.

[46] Racine, J. S., J. Hart and Q. Li (2006). Testing the Significance of Categorical Predictor Variables in Nonparametric Regression Models. *Econometric Reviews* 25 (4), 523-544.

[47] Racine, J.S. and Q. Li (2004). Nonparametric Estimation of Regression Functions with both Categorical and Continuous Data. *Journal of Econometrics* 119 (1), 99–130.

[48] Ray, S.C (1991). Resource Use Efficiency in Public Schools. A Study of Connecticut Data. *Management Science* 37, 1620-1628.

[49] Ruggiero, J. (1996). On the Measurement of Technical Efficiency in the Public Sector. *European Journal of Operational Research* 90, 553–565.

[50] Ruggiero, J. (1998). Non-Discretionary Inputs in Data Envelopment Analysis. *European Journal of Operational Research* 111, 461–469.

[51] Simar, L. and P. Wilson (2007). Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes. *Journal of Econometrics* 136 (1), 31–64.

[52] Sirin, S.R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research* 75 (3), 417–453.

[53] Thanassoulis, E. and M. Portela (2002). School Outcomes: Sharing the Responsibility between Pupil and School. *Education Economics* 10 (2), 183-207.