The Research Efficiency of US Universities:
a Nonparametric Frontier Modelling Approach

Christopher Bruffaerts
SBS-EM, ECARES, Université Libre de Bruxelles

Bram De Rock
SBS-EM, ECARES, Université Libre de Bruxelles

Catherine Dehon
SBS-EM, ECARES, Université Libre de Bruxelles

ECARES working paper 2013-31
The Research Efficiency of US universities: a nonparametric frontier modelling approach

Christopher Bruffaerts* Bram De Rock † Catherine Dehon‡

Abstract

Understanding the factors that may impact how well universities are transforming a set of inputs into research outputs is of great interest for university and public authorities. The goal of this paper is on the one hand to measure the research efficiencies of US universities and on the other hand to study the impact of environmental variables on them. To reach this objective, the latest techniques in nonparametric frontier models are used with both classic and robust methodologies. Focus is in particular devoted to the impact of the institution type (public or private), the teaching load, the degree of collaboration with industrial partners and the degree of international collaborations on the production process associated to research activities. The impact of the size of a university on the way resources are used regarding research activities is also studied.

Keywords: Conditional efficiency measures, Kernel smoothing, Nonparametric frontiers, Research efficiency, Two-stage regression, University rankings.

1 Introduction

The majority of rankings of World University are based on the hypothesis that the stakeholders are mainly interested in the prestige of universities, and hence look at the volume of production. Most rankings of world universities are based solely on research and teaching outputs. But indicators based only on the volume of research and teaching activities of a university are clearly impacted by the size of the institution. For the research part, those rankings take into account variables such as the number of publications, the number of citations, the number of awards, or the number of PhD degrees awarded in a given institution. Academic Ranking of World Universities (ARWU), the Times Higher Education and the QS World University ranking are somehow based on those measures of research excellence measuring the volume of production in opposition

*Université Libre de Bruxelles, ECARES (European Center for Advanced Research in Economics and Statistics). E-mail: cbruffaerts@ulb.ac.be.
†Université Libre de Bruxelles, ECARES. Bram De Rock gratefully acknowledges the European Research Council (ERC) for his starting grant. Bram De Rock is also member of ECORE, the association between CORE and ECARES. E-mail: bderock@ulb.ac.be.
‡Université Libre de Bruxelles, ECARES. Catherine Dehon acknowledges research support from the IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy) and from the ARC contract of the Communauté Française de Belgique. Catherine Dehon is also member of ECORE, the association between CORE and ECARES. E-mail: cdehon@ulb.ac.be.
to the level of productivity \(^1\). But the policy makers might be interested much more in “value for money” and reward good practice.

Moreover, rankings in higher education are closely linked to the topic of evaluation. Increasing efforts in bibliometric evaluation of research have been recently devoted to take into account differences that may subsist between research fields, between countries or even between the quality of different type of publications. In particular, the Scimago and Leiden rankings have analysed the complexity to correctly use bibliometric indicators in assessing the quality of research at universities. They propose some strategies to take into account the biases linked to differences across disciplines or different structures of universities.

One of the dimension of research excellence includes the research efficiency of universities, i.e. the ability of universities to transform efficiently a set of research inputs into outputs (see Farrell (1957)). Instead of focusing mainly on measures of volume or prestige that are used in all of the above rankings, we rather in this paper consider the way in which universities obtain such outputs from their research inputs. It is indeed of great interest to understand the production process that is behind research activities. To reach this objective, the latest techniques in the field of nonparametric frontier models are used in such a way that the research efficiency of universities can be assessed. Compared to most studies done on university rankings, this study takes into account on the one hand the inputs that could have led to research outputs and on the other hand variables that can have an impact on the way universities are transforming their inputs into outputs. This latter type of variable which might impact the production process is usually referred as an environmental variable in nonparametric frontier models.

The focus of this paper relies mainly on the following research questions. First, public and private universities are compared with respect to their research efficiency. In the study of Simar and Zelenyuk (2007) for instance, the authors account for the difference in terms of efficiencies between public and private firms in a particular economic system. Universities being seen as firms by most economists, it seems natural to compare both type of institutions. Second, it is not clear how teaching can affect the way research activities are carried out and this is why we tackle this perplexing problem. This complex relationship between both activities has been extensively studied in Walckiers (2008) through contract designs for academics working on both the production of science and teaching. Third, it is yet unclear how the degree of internationalisation (via the number of international collaborators) affects the production process of research. Does having a large international network imply a higher research efficiency? Last, we consider how the size of a university can impact the way in which universities transform their research inputs into outputs (see Zitt and Filliatreau (2006)). For instance, Bonaccorsi et al. (2006) showed that the size of universities affects the research efficiency in a positive way.

Few studies have been done on the efficiency of universities in terms of research and/or teaching activities by using nonparametric frontier models such as Data Envelopment Analysis proposed by Charnes et al. (1978) (see for instance Katharaki and Katharakis (2010), Worthington (2001) and Worthington and Lee (2008)). Bonaccorsi et al. (2006) studied in particular the efficiency of Italian universities but did not pro-

\(^1\)Some rankings have now introduced some indicators measuring the performance of the university (see for example the variable PCP in ARWU).
vide any statistical procedure to test the impact of environmental variables on efficiencies (only a visual tool). This study is to our knowledge the first that considers the statistical significance of those variables when studying the research efficiencies of universities. By combining different nonparametric tools for the efficiency estimator, we are able to determine the variables that are explaining efficiency differentials among US universities.

When trying to explain efficiencies with environmental variables, a natural idea is to use a two stage procedure in which we first estimate somehow the efficiency of universities and afterwards regress those efficiencies on the set of environmental variables. Many researchers have used this two stage procedure in different contexts by using in particular tobit or probit types of regression (see Simar and Wilson (2007) for a rather exhaustive list of studies). However, as was shown by Simar and Wilson (2007) and Daraio et al. (2010), this two stage procedure is in general not appropriate to understand the impact of environmental variables on the efficiencies. The two-stage procedure indeed relies on a strong separability assumption. In other words, it is required that the environmental variables do not affect the support of the input and output variables. An alternative to the inappropriate use of the two-stage approach is to include directly environmental variables in the computation of the efficiencies through of a probabilistic approach (see Daraio and Simar (2005)). Those conditional efficiencies can afterwards be compared with the unconditional efficiencies and the impact of the environmental variables can be examined. The study of the impact of environmental variables on the efficiency of production units in nonparametric frontier models is now mostly available (see Badin et al. (2012)).

When considering ranking universities, careful analysis to universities that might have an atypical profile with respect to other universities is essential. As shown in Dehon et al. (2010), the use of robust tools such as robust Principal Component Analysis (PCA) in their case can be a way to overcome the problem of outlying universities. Influential points are an even more important issue in the context of nonparametric frontier models as a single outlying university can affect the estimated production frontier and therefore impact the efficiency of all other universities. This is an even more important matter given that bibliometric indicators are usually asymmetrically distributed (see Seglen, 1992), suggesting a more careful analysis to outlying points. This motivates the use of the so-called partial frontiers which can be seen as the robust version of the nonparametric estimated frontier. It could indeed be that some outlying universities are masking the real effect of environmental variables on the production process and therefore mislead the researcher.

In Section 2, the data collection process is explained and the choice of variables is motivated. The methodology to study the impact of mixed type of environmental variables (both discrete and continuous) on the production process is exposed in Section 3. The fourth Section presents the empirical results from the database of 124 US universities and finally Section 5 concludes.

2 Data and variables

The main reason for studying universities in the US is that a rather large sample of universities of a same country is available. Because of this, those universities do have in general the same teaching and research systems. Moreover, publications from US
universities are always in English, which avoids part of the debate we have in Europe on bibliometric research indicators because of language differences mainly. Having a sample of US universities guarantees us of a rather homogeneous sample in comparison to European universities for which the system can be completely different from country to country. Basing the analysis on US universities therefore allows fair comparisons between institutions in terms of their research efficiency.

2.1 Data collection

Given that both research inputs and outputs of US universities are needed, two different sources have been used to collect the data:

- The CWTS Leiden ranking 2013, see Waltman et al. (2012). Rankings of universities are usually based on bibliometric indicators as a proxy for the overall output of scientific research. However, it is clear that those indicators are impacted by intrinsic characteristics such as: the research field, the type of journal where the article is published, the number of coauthors on a publication or even the language in which the article is published. The research group of the Leiden university\(^2\) accounts for all those differences in their ranking methodology. The Leiden ranking measures the scientific performance of 500 major universities worldwide among which 124 US universities. The Leiden database has been created from a data cleansing of the Web of Science database. This is why we believe that the Leiden database is one of the most accurate databases with respect to bibliometric indicators.

- The Integrated Postsecondary Education Data System (IPEDS) data collection managed by the National Center for Education and Statistics (NCES). IPEDS collects data on post secondary education in the US in different fields: institutional characteristics, human and financial resources, enrollments, and degrees awarded. Data from this website has been retrieved from a single data file.

Because the US universities that are part of the Leiden database are also part of the IPEDS database, we have a database of 124 US universities. Among them, one can find the ones that are usually in the top 10 of world university rankings such as Harvard University, MIT, Stanford University, Princeton University or Caltech for instance.

2.2 Choice of variables

To gauge the research efficiency of universities is a difficult matter given that many factors can influence how performant a university is in terms of its research excellence. An important step when performing efficiency analysis is first the choice of inputs and outputs and in a second time the choice of the environmental variables. The main indicators used to assess the research excellence of an institution are usually the number of publications, the number of citations as well as the number of PhD degrees awarded (see Bonaccorsi et al. (2006), Katharaki and Katharakis (2010), Worthington (2001) and Worthington and Lee (2008)). However, those output indicators have had numerous criticism. In particular, researchers argue that the total number of publications from a

\(^2\)http://www.leidenranking.com/
university is not always a fair proxy variable to compare research productivity among universities. Publications can indeed depend on the field in which the article is published, on the type of journal/review where it has been published, on the number of coauthors on a publication or on the language of the publication. For instance, Bao et al. (2010) show the problems related to journal rankings in the field of economics in terms of research productivity. To counter the drawbacks mentioned earlier and avoid part of the criticism, the Leiden Ranking 2013 aims at taking those differences into account. We have considered only publications in core journals which, according to the Leiden Ranking, means that the publication does not fall in one of the following category: the journal does not publish in English or the journal has only a small number of references to other journals in the Web of Science database. Moreover, the so called fractional type of counting of the Leiden Ranking 2013 is considered. This counting method does not give equal weight to collaborative and non-collaborative publications. Less weight will be given to a publication with five coauthors for instance. The fractional type of counting allows fairer comparisons between both universities and fields. More information about those counting methods can be found on the Leiden website.

If output indicators of research excellence can be easily identified, it is not straightforward to determine which inputs have been used to obtain such outputs. We believe however that, as in many other studies (see Bonaccorsi et al. (2006), Katharaki and Katharakis (2010) and Worthington (2001)), input indicators for research can be divided into both human and financial capital. The former is measured on the one hand by the number of professors and associate professors and on the other hand by the number of assistant professors. The reason for separating them into two distinct inputs is that professors and associate professors are expected to have more experience in terms of publications and research activities then assistant professors. For the financial resources, we use the research expenses per Full Time Equivalent (FTE) to assess how much capital has been invested in research activities.

Regarding the environmental variables, there are many variables that might affect the research efficiency of a university. It could be that the type of university (private or public) impacts the research efficiency of US universities. Are public universities that have less financial and human resources less efficient in terms of research compared to private institutions? As was pointed out in Bonaccorsi et al. (2006) and Bonaccorsi et al. (2007), the size of universities can play an important role. The three usual missions of a university being teaching, research activities and collaboration with non-university partners (industry or public service), it seems natural to study the impact of the other two primary activities of a university on its research efficiency. Teaching activities usually take a substantial part of the time of a professor. However, one can ask whether teaching is favorable or detrimental to the way research activities are being carried out. Universities that are part of the database are more research oriented and therefore the teaching load in those institutions remains rather low compared to some other universities. Collaboration with the industry or the public service might as well impact research activities. Are collaborations more time consuming or are they on the opposite more stimulating and promoting research? Beside the teaching and the industrial sector activities of a university, we also consider the degree to which a university is collaborating with other universities for research purposes. Does having many international collaborators imply having a higher research efficiency? Another important factor that has
to be taken into account is the quality of the research being conducted or equivalently the quality of the journal/review where the research article is published. Writing a high quality paper or publishing in a top journal can also be more time consuming. This motivates the inclusion of a variable measuring the global quality of publications of a university. To do so, we use the proportion of articles that were published in the top 10% journals according to the criteria used in the Leiden Ranking. The average salary of professors in one university is used to see whether it can be an incentive for being more efficient in terms of research. Indeed, universities proposing high salaries for professorships can attract top researchers. To have an idea of the extent to which human and social sciences are studied in a university, we consider the proportion of first year students enrolled in those types of fields. This measure is a proxy to the importance of human and social sciences in an institution. All input and output indicators as well as the environmental variables are summarized in Table 1. From summary statistics given in Table 2, it can be noted the existing heterogeneity among US universities for all input and output indicators as well as environmental variables. For instance, the ratio between the maximum and the minimum value for the research expenses is 190, for the number of publications 19 and for the number of PhD degrees 46.

Table 1: Different variables to assess the research efficiency of US universities.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
<th>Year</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input indicators</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expenses</td>
<td>Average Research expenses per FTE (dollars)</td>
<td>2006-2007</td>
<td>IPEDS</td>
</tr>
<tr>
<td>Associate professors</td>
<td>Number of full time professors and associate professors</td>
<td>2007</td>
<td>IPEDS</td>
</tr>
<tr>
<td>Assistant professors</td>
<td>Number of full time associate professors</td>
<td>2007</td>
<td>IPEDS</td>
</tr>
<tr>
<td><strong>Output indicators</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publications</td>
<td>Total number of Publications</td>
<td>2008-2011</td>
<td>Leiden</td>
</tr>
<tr>
<td>Citations</td>
<td>Total number of Citations</td>
<td>2008-2011</td>
<td>Leiden</td>
</tr>
<tr>
<td>PhD degrees</td>
<td>Total number of Phd degrees awarded</td>
<td>2010-2011</td>
<td>IPEDS</td>
</tr>
<tr>
<td><strong>Environmental</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>Public or private institution</td>
<td>-</td>
<td>IPEDS</td>
</tr>
<tr>
<td>Size</td>
<td>Institution size category (total number of students enrolled)</td>
<td>2008</td>
<td>IPEDS</td>
</tr>
<tr>
<td>Teaching Load</td>
<td>Teacher/student ratio</td>
<td>2008</td>
<td>IPEDS</td>
</tr>
<tr>
<td>Industry</td>
<td>Proportion of collaborative publications with industry</td>
<td>2008-2011</td>
<td>Leiden</td>
</tr>
<tr>
<td>International</td>
<td>Proportion of international collaborative publications</td>
<td>2008-2011</td>
<td>Leiden</td>
</tr>
<tr>
<td>Interinstitutional</td>
<td>Proportion of interinstitutional collaborative publications</td>
<td>2008-2011</td>
<td>Leiden</td>
</tr>
<tr>
<td>Salary</td>
<td>Average salary of professors</td>
<td>2006-2008</td>
<td>IPEDS</td>
</tr>
<tr>
<td>Top publications</td>
<td>Proportion of top 10% publications</td>
<td>2008-2011</td>
<td>Leiden</td>
</tr>
<tr>
<td>Percentage Human/social</td>
<td>Proportion of students in Human/Social sciences</td>
<td>2008</td>
<td>IPEDS</td>
</tr>
</tbody>
</table>

For the variable “Size”, we account 22 small scale universities (less than 10,000 students), 23 medium scale universities (between 10,000 and 20,000 students) and 79 large scale universities (more than 10,000 students). The database includes 82 public
Table 2: Descriptive statistics of the sample of $n = 124$ US universities.

<table>
<thead>
<tr>
<th>Variable</th>
<th>mean</th>
<th>median</th>
<th>min</th>
<th>max</th>
<th>std dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input indicators</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expenses</td>
<td>23265</td>
<td>9053</td>
<td>1616</td>
<td>307486</td>
<td>41010</td>
</tr>
<tr>
<td>Associate professors</td>
<td>991.5</td>
<td>939</td>
<td>135</td>
<td>2170</td>
<td>446.1</td>
</tr>
<tr>
<td>Assistant professors</td>
<td>503.4</td>
<td>445.5</td>
<td>48</td>
<td>1569</td>
<td>282.3</td>
</tr>
<tr>
<td><strong>Output indicators</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publications</td>
<td>5486</td>
<td>4476</td>
<td>1560</td>
<td>29812</td>
<td>3927.7</td>
</tr>
<tr>
<td>Citations</td>
<td>7314</td>
<td>4998</td>
<td>1462</td>
<td>53624</td>
<td>6687.5</td>
</tr>
<tr>
<td>PhD degrees</td>
<td>638.1</td>
<td>543</td>
<td>39</td>
<td>1811</td>
<td>417.1</td>
</tr>
<tr>
<td><strong>Environmental</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teaching Load</td>
<td>14.9</td>
<td>16</td>
<td>1</td>
<td>30</td>
<td>5.9</td>
</tr>
<tr>
<td>Industry</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>11</td>
<td>1.5</td>
</tr>
<tr>
<td>International</td>
<td>31</td>
<td>30</td>
<td>22</td>
<td>50</td>
<td>5.1</td>
</tr>
<tr>
<td>Interinstitutional</td>
<td>72</td>
<td>72</td>
<td>65</td>
<td>81</td>
<td>3.32</td>
</tr>
<tr>
<td>Salary</td>
<td>126569</td>
<td>124636</td>
<td>89237</td>
<td>191703</td>
<td>21481</td>
</tr>
<tr>
<td>Top publications</td>
<td>13.44</td>
<td>13</td>
<td>7</td>
<td>25</td>
<td>3.6</td>
</tr>
<tr>
<td>Percentage Human/social</td>
<td>40</td>
<td>44</td>
<td>0</td>
<td>81</td>
<td>20.2</td>
</tr>
</tbody>
</table>

Universities and 42 private ones. For the variable “Salary”, we consider in the sequel the logarithm of this variable because of a strong asymmetry.

2.3 Choice of orientation

In this study, to assess how far universities lie from the production frontier, we use the hyperbolic type of orientation that was originally proposed by Färe et al. (1985). As stressed by Wheelock and Wilson (2008), there are many advantages for using this specific type of distance in our context. First, no a priori orientation has to be chosen (such as the input, output or the directional orientation) given that the hyperbolic orientation simultaneously adjusts both inputs and outputs. The use of this type of distance, in the context of universities can of course be criticized but we believe that it is in this case a more adequate indicator of efficiency than the other orientations. It is indeed not obvious whether university authorities have the control only on inputs or only on outputs. For instance, Abbott and Doucouliagos (2003) study the efficiency of Australian universities by considering the input orientation while Johnes and Yu (2008) study the efficiency of Chinese universities following an output orientation. Second, a major advantage of the hyperbolic orientation is that a university is only compared to its dominating set (i.e. universities that are using less inputs and at the same time producing more outputs than the university being analyzed). This is for instance not the case for the input and output orientations for which a small scale university can be compared with a much larger one. Moreover, since our motivation is to get a global picture of the research efficiency of US universities, we believe the choice of the hyperbolic orientation to be in this case appropriate.
3 Methodology

3.1 Probabilistic formulation and partial frontiers

It was shown by Cazals et al. (2002) that the production set which is defined as \( \Psi = \{(x, y) \in \mathbb{R}_{+}^{p+q} \mid x \text{ can produce } y \} \) can be reformulated using a probabilistic approach. Not only has this idea enabled to introduce partial frontiers (robust versions of the complete production frontier, see Aragon et al. (2005)) but it has also been used to define conditional efficiencies (see Daraio et al. (2005)). The production set can indeed be completely defined by means of the following joint distribution function:

\[
H_{XY}(x, y) = P[X \leq x, Y \geq y],
\]

which represents the probability that a unit performing at level \((x, y)\) is being dominated. The partial production frontier of order \(\alpha\) (\(\alpha \in [0, 1]\)) can be defined as follows:

\[
\Psi_{\alpha} = \{(x, y) \in \mathbb{R}_{+}^{p+q} \mid H_{XY}(x, y) > 1 - \alpha\}.
\]

If \(\alpha = 1\), one can recover the complete production set: \(\Psi_{\alpha} = \Psi\). The hyperbolic type of estimator of order \(\alpha\) is defined as:

\[
\gamma_{\alpha}(x, y) = \sup\{\gamma > 0 \mid H_{XY}(\gamma^{-1}x, \gamma y) > 1 - \alpha\},
\]

where the traditional hyperbolic type of estimator for \((x, y)\) is \(\gamma(x, y) = \gamma_1(x, y)\).

The nonparametric estimation of the distribution \(H_{XY}(x, y)\) is given as:

\[
\hat{H}_{XY,n}(x, y) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x, Y_i \geq y),
\]

and yields the following order-\(\alpha\) hyperbolic type of estimator:

\[
\hat{\gamma}_{\alpha,n}(x, y) = \sup\{\gamma > 0 \mid \hat{H}_{XY,n}(\gamma^{-1}x, \gamma y) > 1 - \alpha\}.
\]

This nonparametric hyperbolic efficiency measurement gives the amount by which inputs and outputs must be respectively reduced and increased in order for the DMU under analysis to be dominated by \((1 - \alpha) \times 100\%\) production units. This type of estimator has been extensively studied by Wheelock and Wilson (2008) and Bruyraerts et al. (2013). It is asymptotically normal and possess good robustness properties for an adequate choice of \(\alpha\). For a value of \(\alpha = 1\), the usual Free Disposal Hull type of estimator for the hyperbolic orientation is recovered (see Deprins et al. (1984)).

3.2 A conditional efficiency analysis

At first, conditional efficiencies were studied only for continuous type of environmental variables. Recently, techniques to include both discrete nominal (unordered) and discrete ordered environmental variables in efficiency analysis have emerged. In particular, De Witte and Kortelainen (2013) have adapted to the setting of conditional efficiency the ideas from Li and Racine (2008) for estimating nonparametrically a conditional cumulative distribution function using mixed categorical and continuous variables. Given
the above probabilistic formulation of the production set, one can consider the following joint conditional distribution:

\[ H_{XY|Z}(x, y | z) = P[X \leq x, Y \geq y | Z = z]. \]

The nonparametric estimation of this conditional distribution is done through smoothing techniques. The above distribution can be estimated as follows:

\[
\hat{H}_{XY|Z,n}(x, y | z) = \frac{\sum_{i=1}^{n} \mathbb{1}(X_i \leq x, Y_i \geq y | Z = z) K_h(z, Z_i)}{\sum_{i=1}^{n} K_h(z, Z_i)},
\]

where \( K_h(z, Z_i) \) is a generalized kernel. Depending on the type of variables included in the vector of environmental variables, this type of kernel has to be adapted. For a given vector of observed environmental variables \( z_i = (z_{1i}, z_{2i}, z_{3i}) \) containing respectively \( r \) continuous variables, \( v \) ordered variables, \( w \) unordered variables and \((h_{1c}, h_{2c}, h_{3c})\) their associated bandwidths, its general form is given as follows:

\[
K_h(z, z_i) = r \prod_{s=1}^{r} \frac{1}{h_{cs}} f^c \left( \frac{z_{cs} - z_{is}}{h_{cs}} \right) \prod_{s=r+1}^{r+v} l^o(\delta_{is}, h_{os}) \prod_{s=r+v+1}^{r+v+w} l^u(\delta_{is}, h_{us}),
\]

where \( f^c, l^o, \) and \( l^u \) are respectively kernels for the continuous, the discrete ordered and the discrete unordered variables. Aitchison and Aitken (1976) proposed the following discrete kernel for unordered variables:

\[
l^u(\delta_{is}, \delta_{is}, h_{us}) = \begin{cases} 1 - h_{us} & \text{if } \delta_{is} = \delta_{is} \\ \frac{h_{us}}{(\delta_{is} - 1)} & \text{if } \delta_{is} \neq \delta_{is}. \end{cases}
\]

For the ordered variables, the kernel of Li and Racine (2007) is used:

\[
l^o(\delta_{is}, \delta_{is}, h_{os}) = (h_{os})^{\delta_{is} - \delta_{is}^{-1}}.
\]

Many kernels can be chosen for the continuous type of variable but we choose to use the usual Epaneschnikov kernel.

The conditional order \( \alpha \) hyperbolic efficiency estimator can therefore be defined as:

\[
\hat{\gamma}_{\alpha,n}(x, y | z) = \sup \{ \gamma > 0 | \hat{H}_{XY|Z,n}(\gamma^{-1}x, \gamma y | z) > 1 - \alpha \}.
\]

### 3.3 Data-driven bandwidth selection

When dealing with kernel based methods, the choice of the bandwidths is crucial for both estimation and inference. The idea originally proposed by Badin et al. (2010) is to use a cross-validation methodology that is adapted to the problem of conditional efficiency estimation. The technique is based on Hall et al. (2004) in which the authors use cross-validation to estimate conditional probability densities. Adapting ideas of Hall et al. (2004) and De Witte and Kortelainen (2013) to the hyperbolic case, we consider the following conditional joint distribution:

\[
g(x, y | Z = z) = \frac{f(x, y, Z = z)}{m(Z = z)},
\]
where \( f(x, y, Z = z) \) is the joint density of \((x, y, z)\) and \(m\) is the marginal density of \(Z\). This conditional joint density can be estimated by using the following nonparametric density estimates:

\[
\hat{f}(x, y, z) = \frac{1}{n} \sum_{i=1}^{n} K_h(z, z_i) J_{h_x}(x, x_i) J_{h_y}(y, y_i),
\]

\[
\hat{m}(z) = \frac{1}{n} \sum_{i=1}^{n} K_h(z, z_i),
\]

where \( J_{h_x}(x, x_i) \) and \( L_{h_y}(y, y_i) \) represent the multivariate kernels for \(x\) and \(y\) respectively. They are defined as

\[
\prod_{j=1}^{p} \frac{1}{\pi_x} l\left(\frac{x_j - x_{ij}}{\pi_x}ight) \quad \text{and} \quad \prod_{j=1}^{q} \frac{1}{\pi_y} l\left(\frac{y_j - y_{ij}}{\pi_y}\right)
\]

with \( l(.) \) a univariate kernel that has been previously defined. The idea of Hall et al. (2004) is to, via cross-validation techniques, to find the bandwidths that minimize the discrepancy between both \( \hat{g}(\cdot | \cdot) \) and \( \hat{g}(\cdot | \cdot) \). Some adjustments for the bandwidths need to be done given that it is a conditional cdf that is being estimated and not a pdf.

For the hyperbolic case, estimating the conditional joint distribution \( g(\cdot | \cdot) \) requires smoothing in both inputs and outputs. This is a main difference with both the input and output orientation where smoothing is needed on outputs and inputs respectively. This cross-validation procedure will yield bandwidths \((h_x, h_y, h^c, h^n, h^u)\) although the bandwidths \((h_x, h_y)\) are not used in the estimation of the conditional efficiency. It was shown in Badin et al. (2010) and De Witte and Kortelainen (2013) by means of simulations and under different settings that this data-driven bandwidth selection works well.

### 3.4 Detecting the impact of the environmental variables

In order to detect whether environmental variables impact the production process, the ratio of conditional and unconditional efficiencies is regressed nonparametrically on the set of environmental variables. By doing so, Badin et al. (2012) show how the impact of environmental factors can be detected and measured. Most of the literature has focused on two-stage approaches to explain the way in which those exogeneous variables can affect the efficiencies of production units. Although it has been already widely used, Simar and Wilson (2007) show that this approach is inappropriate whenever the so called separability condition is not fulfilled. Measuring the impact of environmental variables on the production process relies on the analysis of the ratio of conditional and unconditional hyperbolic efficiencies:

\[
\hat{R}_H(x, y | z) = \frac{\hat{\gamma}_n(x, y | z)}{\hat{\gamma}(x, y)},
\]

or to its robust version:

\[
\hat{R}_{H,\alpha}(x, y | z) = \frac{\hat{\gamma}_n(x, y | z)}{\hat{\gamma}_{\alpha,n}(x, y)}.
\]

In a first attempt to understand the way in which the environmental variables can impact the production process, one could simply regress linearly the above ratios on the set of environmental variables. Although linear, this regression could already give an
idea to the researcher whether in average effects are positive, negative or neutral. To capture nonlinear effects, Daraio and Simar (2005) propose to use for instance a smooth nonparametric kernel regression:

$$Q^*_i = f(z_i) + \varepsilon_i \quad \forall i = 1, \ldots, n$$

where $Q^*_i$ is either the classic or the robust efficiency ratio, with $\varepsilon_i \sim N(0,1)$ and with $z_i = (z^c_i, z^o_i, z^u_i)$ being the vector containing the continuous, the discrete ordered and the discrete unordered variables respectively.

It was shown in Daraio and Simar (2005) how the impact of single continuous environmental variables can be detected visually from this nonparametric regression. In the case of the hyperbolic efficiency, a horizontal fit implies that the continuous environmental variable has no impact on the production process whereas an increasing (decreasing) curve suggests that it has a favorable (detrimental) effect on it. When dealing with more than one environmental variable, Daraio and Simar (2007) have suggested looking at the marginal effects of one variable at a time (defined as partial regression plots). In other words, the authors study the impact on the efficiency ratio when only one variable is able to change and all other variables are kept constant. As in De Witte and Kortelainen (2013), it is possible to examine the same effect but for different fixed values of the environmental variables (such as the median or quartiles) in order to detect different behaviors among production units. This methodology however is purely based on the visualization of the nonparametric curve. Badin et al. (2012), by means of subsampling techniques, provide a way of creating confidence intervals for the efficiency ratio. To test the statistical significance of the impact of $Z$ on the production process, we follow the approach of Racine and Li (2004) in which they use a local linear methodology to estimate nonparametrically the regression curve by taking into account a mixture of both continuous and discrete type of variables. The idea of their method is to minimize the following objective function:

$$\min_{\alpha, \beta} \sum_{i=1}^{n} \left( Q^*_i - \alpha - (z^c_i - z^c) \beta \right)^2 K_h(z, z_i),$$

where as before $K_h(z, z_i)$ is the generalized kernel.

Nonparametric tests, which can be regarded as equivalent of standard t-tests for OLS regressions, were proposed in Racine et al. (2006) to test the significance of both continuous and discrete variables. Those tests are based on bootstrapping the residuals of the nonparametric regression defined above. They allow to see whether a variable of any type has a significant impact on the efficiency ratio.

4 Empirical Results

4.1 Conditional and unconditional efficiencies

The above methodology is applied to the database of US universities that was presented in Section 2. Some outlying universities might however have undesirable effects on both the unconditional and conditional efficiency of other universities. Because of this, we use both the classic ($\alpha = 1$) and the robust estimators. The choice of the value of $\alpha$ for
the robust type of estimator is still a tedious matter. However, some simple methods to determine appropriate values have been proposed by Daouia and Gijbels (2011) and Bruffaerts et al. (2013) by looking at the evolution of efficiency estimates as a function of the parameter $\alpha$ and identifying jumps on this plot. In our case, the value is set to $\alpha = 0.975$ as can be seen from plots on Figure 1. The first figure on the left hand side depicts the evolution of the average efficiencies as a function of $\alpha$ while the second figure shows the boxplots of efficiencies for different values of $\alpha$. As can be noticed, jumps and changes in the efficiencies occur at the value $\alpha = 0.975$, which results typically in $124 \times (1 - 0.975) \approx 3$ points being outside of the frontier for each university being analysed. By leaving out only 3 universities will allow to perform a robust analysis while at the same time keeping the estimated partial type of frontier close to the real production frontier. It is important to consider the robust type of estimator because it can indeed be that for the classic efficiency ratio some outlying universities hide the real effect of the environmental variables.

Figure 1: Evolution of the average efficiencies (left plot) and boxplots of the research efficiencies of US universities for different values of $\alpha$ (right plot).

4.2 Impact of the environmental variables

Once $\alpha$ has been chosen, the impact of all environmental variables (continuous and discrete) can be studied in both the classical and robust way. First, the impact of the discrete type of variable which is the type of university (public or private) is studied. As can be seen from Figure 2, public universities are in average more efficient regarding their research than private institutions. To know whether this “visual” effect is statistically
significant, the Hall et al. (2004) bootstrap methodology is performed and results are presented in Table 3. The variable is highly significant (with both classical and robust estimators), which means that the type of institution is an important factor that can explain efficiency differentials among universities.

Figure 2: Average impact of the Type (Public/Private) of US universities on their research efficiency.

<table>
<thead>
<tr>
<th>Type</th>
<th>efficiency_ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public</td>
<td>0.82</td>
</tr>
<tr>
<td>Private</td>
<td>0.84</td>
</tr>
<tr>
<td>Public</td>
<td>0.86</td>
</tr>
<tr>
<td>Private</td>
<td>0.88</td>
</tr>
<tr>
<td>Public</td>
<td>0.90</td>
</tr>
<tr>
<td>Private</td>
<td>0.92</td>
</tr>
<tr>
<td>Public</td>
<td>0.94</td>
</tr>
<tr>
<td>Private</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Table 3: P-values for the impact of the Type of institution on the research efficiency for both classic ($\alpha = 1$) and robust ($\alpha = 0.975$) methodologies.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Classic</td>
</tr>
<tr>
<td>Type (Public/Private)</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.95</td>
</tr>
</tbody>
</table>

As for the discrete type of variable, the impact of the continuous type of environmental variables is studied. Table 4 shows the different settings that have been considered. In the first model, one can learn that the teaching load is highly significant whereas collaboration with industrial partners is not. Figure 3 shows the type of impact the teaching load has on the research efficiency for the classic and robust efficiency ratios. Clearly, teaching has a favorable impact given the increasing regression line although the effect seems to become neutral beyond a certain point. To better understand how teaching and collaboration with the industry impact the research efficiency, we show in Figure 3 the joint impact of those two “competing” academic activities on the efficiency ratio.

From the second model of Table 4, we learn that with the classic ratios neither international nor interinstitutional collaborations impact the research efficiency for the classic ratio. It seems however that some outlying universities are masking the real
Table 4: P-values for the impact of the continuous environmental variables on the research efficiency for different models and for both classic ($\alpha = 1$) and robust ($\alpha = 0.975$) methodologies.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Classic</td>
<td>Robust</td>
<td>Classic</td>
</tr>
<tr>
<td>Teaching Load</td>
<td>0.001***</td>
<td>&lt;0.001***</td>
<td></td>
</tr>
<tr>
<td>Industry</td>
<td>0.24</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>International</td>
<td>0.98</td>
<td>0.01***</td>
<td></td>
</tr>
<tr>
<td>Interinstitutional</td>
<td>1</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>Log Salary</td>
<td></td>
<td>0.06*</td>
<td>0.11</td>
</tr>
<tr>
<td>Top Pub.</td>
<td></td>
<td>0.99</td>
<td>0.36</td>
</tr>
<tr>
<td>% Human/Social</td>
<td></td>
<td>0.84</td>
<td>0.07*</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.53</td>
<td>0.69</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Figure 3: Impact of the teaching load on the research efficiency (classic on the left and robust on the right).

effect of the international collaborations given that it is highly significant for the robust efficiency ratio. Figure 5 shows that its impact follows an inverted “U”-shaped curve (the robust is less shaped than the classic) which means that international collaborations are favorable only up to a certain point, after that the effect becomes detrimental. The robust joint impact of international and interinstitutional collaborations is shown in Figure 6.
Figure 4: Joint impact of the teaching load and the collaboration with industry on the research efficiency.

Figure 5: Impact of international collaborations on the research efficiency (classic on the left and robust on the right).

Finally, the last model of Table 4 shows that for the classic ratios the salary has a significant favorable impact on the efficiency (when looking at partial regression plots). This significant effect however disappears once we account for potential outlying universities. It can be noticed that the quality of publications has no impact at all on the research efficiency. Interestingly, the percentage of human and social sciences impacts in a favorable way the robust efficiency ratios.

Finally, Table 5 includes the p-values when combining the type of the institution.
Figure 6: Joint impact of international and interinstitution collaborations on the research efficiency (robust).

and all other continuous type of variables. As pointed out previously, being a public or a private institution is an important factor in explaining efficiencies among US universities. The teaching load and the collaboration with the industry remain important factors for both the classic and robust efficiency ratios. For the type of collaboration, only interinstitutional collaborations have an impact on the production process. As will be shown later, the average effect of having many interinstitutional collaborations is detrimental. The degree to which a university focuses on human and social sciences has a favorable impact on how universities transform their inputs into outputs.

To get a global overview of the impact of all environmental variables, Figure 7 shows the partial regression plots for each variable. As discussed previously, public institutions are in average much more efficient than private one. From this plot, one can learn that as the teaching activities increase, the efficiency linked to research activities also increases. The teaching load has therefore a positive impact on the research efficiency of universities. It is important to note that only the top US universities have been considered in this study and this is why the teaching load is not so large in comparison to other US universities that are not in the database. Beyond a certain teaching load, the increasing line is expected to become more flat at a certain point. Collaboration with the industry has a favorable impact but as for the teaching the impact is expected to become flatter beyond a certain degree of collaboration. Both the salary and the quality of publications are not influencing the research efficiency meaning in particular that the average salary does not seem to be an incentive to become more efficient in terms of research. The percentage of human and social sciences in a university has clearly a positive impact on the way inputs are transformed into research outputs. One explanation for this is that the costs related to research activities related to human and social sciences are much lower than for the hard sciences or the medical sciences. To know whether this is a plausible explanation, the input indicator related to the
Table 5: P-values for the impact of both discrete and continuous environmental variables on the research efficiency for both classic ($\alpha = 1$) and robust ($\alpha = 0.975$) methodologies.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Classic</td>
</tr>
<tr>
<td>Type</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Teaching Load</td>
<td>0.028**</td>
</tr>
<tr>
<td>Industry</td>
<td>0.036**</td>
</tr>
<tr>
<td>International</td>
<td>0.252</td>
</tr>
<tr>
<td>Interinstitutional</td>
<td>0.062*</td>
</tr>
<tr>
<td>Log Salary</td>
<td>0.378</td>
</tr>
<tr>
<td>Top Pub.</td>
<td>0.288</td>
</tr>
<tr>
<td>% Human/Social</td>
<td>0.042**</td>
</tr>
</tbody>
</table>

$R^2$ 0.55 0.71

research expenses was removed and the above methodology was redone. For the classic procedure, the percentage of human and social sciences has no longer a significant impact on the research efficiency (p-value=0.25). For the robust case however, the variable becomes significant again (p-value=0.05) but the type of impact passed from favorable to detrimental (decreasing line for the ratio of robust efficiencies). When removing the research expenses as an input indicator confirms the above explanation that human and social sciences are less costly compared to other types of sciences.

In order to show how misleading a two-stage approach could be in this case, Table 6 shows the differences between the two-stage regression and both the parametric (linear) and nonparametric conditional methodologies presented before. In this Table, we show on the one hand the results of the nonparametric conditional analysis for which the type of impact is shown with a “+” or “−” as well as its significancy and on the other hand the coefficients of the two regression procedures. As can be seen, the type of impact (either positive or negative) is generally the same between the three procedures. However, it is not the case for the significancy of the variables. For the three types of analysis, the institution type remains always highly significant. The teaching load is significant except for the linear regression on the conditional efficiencies. This could be explained by the fact that the impact of teaching activities has locally a more nonlinear type of impact on the research efficiency (see Figure 3). The two stage procedure would have captured a significant effect for the collaboration with industrial partners. Contrarily to the significant impact of the salary in the traditional regression, the salary does not affect in average how research resources are being used. This Table suggests that the separability condition is in this case probably not fulfilled, which means that those environmental variables impact the production set related to US universities. Moreover, the nonparametric conditional analysis shows to be more flexible than the usual linear regression of the conditional efficiency ratios.
Finally, we look at the managerial efficiency of US universities with respect to the size of the universities. As explained in Badin et al. (2010), the residuals from the regressions (either parametric or nonparametric) with the conditional efficiency estimator can be seen as a measure of the managerial efficiency of production units. This type of efficiency somehow represents a whitening of the conditional efficiencies from the effects due to the environmental variables. In Figure 8, boxplots of the managerial efficiencies for the different sizes of universities are depicted. As can be noticed, the size of a university does not seem to impact much the way universities are using their resources in terms of research. An ANOVA test was performed and confirms the fact that we do not have enough evidence to say that the size of a university matters when controlling for the other variables that have been introduced previously. As it has been shown in previous studies, university rankings measure the prestige and volume linked to research activities and do crucially depend on the size of the institutions that are part of the ranking. In this study we show that with respect to the research efficiency there is no size effect once we control...
Table 6: Differences between the nonparametric conditional analysis and the two-stage regression with all environmental variables for the classic and robust efficiency ratios.

<table>
<thead>
<tr>
<th>Variable</th>
<th>NP reg.</th>
<th>Linear reg.</th>
<th>Two-stage reg.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Classic</td>
<td>Robust</td>
<td>Classic</td>
</tr>
<tr>
<td>Type (Public bench.)</td>
<td>***</td>
<td>***</td>
<td>-0.34***</td>
</tr>
<tr>
<td>Teaching Load</td>
<td>+**</td>
<td>+</td>
<td>0.005</td>
</tr>
<tr>
<td>Industry</td>
<td>+**</td>
<td>+</td>
<td>0.0026***</td>
</tr>
<tr>
<td>International</td>
<td>-</td>
<td>-</td>
<td>-0.003</td>
</tr>
<tr>
<td>Interinstitutional</td>
<td>-*</td>
<td>-</td>
<td>-0.01**</td>
</tr>
<tr>
<td>Log Salary</td>
<td>+</td>
<td>+</td>
<td>0.2</td>
</tr>
<tr>
<td>Top Pub.</td>
<td>-</td>
<td>-</td>
<td>0.01**</td>
</tr>
<tr>
<td>% Human/Social</td>
<td>+**</td>
<td>+***</td>
<td>0.004***</td>
</tr>
</tbody>
</table>

for other variables. The implication of such result is that when financing universities, public authorities should therefore focus attention on research efficiency rather than on rankings.

Figure 8: Boxplots of the managerial efficiencies of universities by their Size (from Model 4 with robust ratios).
5 Conclusion

In this paper, we study the factors that may impact the research efficiency of US universities. To this purpose, the latest techniques in nonparametric frontier models have been applied on a database of 124 US universities. This study has enabled us to understand the way in which environmental variables affect the production process regarding research activities. From the empirical results, it is clear that the type of university plays an important role: public universities are more efficient than private ones in terms of their research. Public authorities which are financing public universities can be confident that resources are being well used in terms of research. This sends a clear message to authorities in that it is important for science to keep public universities. Teaching activities and collaboration with industrial partners are positive to the research efficiency although we believe that the effect will become flatter and could become detrimental as the share of teaching load and collaboration with the industry increases. Interinstitutional collaborations have a significant detrimental impact on the way research activities are carried out. This can be explained by the fact that in top universities researchers tend not to collaborate with colleagues but rather target an international scope when it comes to publications even if the degree of internationalization is not significant in our study. Contrarily to previous studies, the size of a university itself has no impact on the way both research inputs and outputs are managed when we measure the research efficiency and control for a set of other variables. This study has also showed that the researcher can be misled first by an inadequate methodological approach (two-stage approach) or even by influential universities. Both the nonparametric conditional efficiency as well as the robust approach has been applied to assess accurately the research efficiency of US universities. This type of analysis could have some policy implications for authorities of universities.
References


