ISSN 2281-4299

DIPARTIMENTO DI INGEGNERIA INFORMATICA
AUTOMATICA E GESTIONALE ANTONIO RUBERTI

SAPIENZA
UNIVERSITÀ DI ROMA

Rankings and university performance: a conditional multidimensional approach

Cinzia Daraio
Andrea Bonaccorsi
Léopold Simar

Technical Report n. 9, 2014

Rankings and university performance: a conditional multidimensional approach

Cinzia Daraio* Andrea Bonaccorsi Léopold Simar

June 17, 2014

Abstract:

University rankings are the subject of a paradox: the more they are criticized by social scientists and experts on methodological grounds, the more they receive attention in policy making and the media. In this paper we attempt to give a contribution to the birth of a new generation of rankings, one that might improve on the current state of the art, by integrating new kind of information and using new ranking techniques. Our approach tries to overcome four main criticisms of university rankings, namely: monodimensionality; statistical robustness; dependence on university size and subject mix; lack of consideration of the input-output structure. We provide an illustration on European universities and conclude by pointing on the importance of investing in data integration and open data at European level both for research and for policy making.

Keywords: Rankings, European universities, DEA, conditional directional distances, robust frontiers, bootstrap

*__Daraio__: Department of Computer, Control and Management Engineering Antonio Ruberti (DIAG), University of Rome "La Sapienza", Via Ariosto, 25 Rome - 00185 Italy; email daraio@dis.uniroma1.it. Research supported by Sapienza University of Rome, "Progetto di Ateneo 2013". **Bonaccorsi**: on leave, Department of Electrical Systems and Automation, University of Pisa, Italy; email a.bonaccorsi@gmail.com. **Simar**: ISBA, Université Catholique de Louvain, Louvain-la-Neuve, Belgium and DIAG University of Rome "La Sapienza"; email leopold.simar@uclouvain.be. Research supported by IAP Research Network P7/06 of the Belgian State (Belgian Science Policy).

1 Introduction and research questions

University rankings are the subject of a paradox: the more they are criticized by social scientists and experts on methodological grounds, the more they receive attention in policy making and the media. Rather than adding to the large literature on the methodological shortcomings of the existing rankings, this paper tries to give a contribution to the birth of a new generation of rankings, one that might improve on the current state of the art both in substantive and methodological bases. We provide two contributions: integrating new kind of information and using new ranking techniques.

The main criticisms (that we report in their historical order of introduction in the literature) addressed to university rankings, which we examine in detail in Section 2, can be summarized as follows:
(a) Monodimensionality
(b) Statistical robustness
(c) Dependence on university size and subject mix
(d) Lack of consideration of the input-output structure.
According to several authors, world rankings suffer from focusing only on the research dimension, which is more visible and easier to measure using external observations. A call for integrating the existing rankings with the educational perspective is in order. Yet several studies call into question the statistical properties of the rankings, irrespective of their substantive content, while others show that rankings systematically distort the representation in favour of large and established universities, and of universities in which scientific and technological disciplines, with particular reference to medical disciplines, are dominant. Finally, a few authors have raised the issue of whether it is acceptable to rank universities worldwide, without any consideration of the differences in resources made available to them by their respective national governments, or their input-output structure.

In this paper we provide an experiment that addresses all these issues, with reference to universities in Europe. The experiment might be replicated in USA and in several Asian countries, which have data comparable to the ones we use here.

First, we reduce monodimensionality by integrating data on research output (basically, scientific publications) with data on the teaching mission of universities. This is a major departure from existing rankings. The integration has been made possible by the creation of the Eumida census of Higher Education Institutions in Europe, a project supported by the European Commission and Eurostat. In addition, we use data that refer to the quality of research. Thus by integrating data on education and research, and by including data not only on students but on degrees, we address the monodimensionality issue. In future studies other indicators (not available for this study) might be included, such as third mission, regional engagement and research infrastructures, leading to even more comprehensive analyses.

Second, we propose a ranking technique that is based on estimators that are robust to extreme values and outliers (as illustrated in Section 4) and delivers confidence intervals for the estimates (as illustrated in Appendix A), allowing the analyst to fully understand the statistical properties of the ranking score we propose.

Third, we address the dependence of rankings on size and subject mix by using a novel technique, called directional conditional efficiency analysis. As illustrated in the methodological section, this technique permits the estimation of efficiency measures net of the impact of size of universities (as proxied by the number of students) and net of the subject mix. This is another major departure from existing rankings. While our data do not allow any estimation in the fields of Humanities and Social Sciences (HSS), due to the limitations of current databases, for the first time we consider the subject mix of universities, as proxied by the specialization index of universities.

Fourth, the ranking we propose is based on an explicit input-output structure. We take benefit from the data in the Eumida dataset, that include academic and nonacademic staff and personnel and non-personnel expenditures, to compute technical efficiency indicators in a multi-input multi-output framework. In this framework a university ranked high is one that makes the best possible use of its resources, on which it may have little discretionary power.

A consolidated literature has applied Data Envelopment Analysis (DEA) in the education sector (see e.g. Sarrico and Dyson, 2000; Sarrico et al. 2009 and Grosskopf et al. 2014 and the references cited therein).

From a methodological point of view, this paper implements in the context of universities rankings the conditional directional distance approach by Daraio and Simar (2014) extending it to derive confidence bounds on the "managerial" efficiency scores robustly estimated. Indeed, as rightly emphasized by Grosskopf et al. (2014, p. 24): "Policy makers are interested in using efficiency scores [...] so it is crucially important to strengthen existing strategies for generating confidence bands around efficiency scores [...]".

Recently, Bonaccorsi, Daraio and Simar (2014) propose a robust directional distance approach to analyse economies of scale and scope in European universities and find that both size and specialization have a statistical significant effect on the efficiency. In this paper we make a step further and estimate the efficiency in the production of research quality (a factor built taking into account international collaborations, normalized impact of research, high quality publications and excellence rate), but considering also the level of teaching and the research output (volume of scientific production) realized, that is considering the level of teaching and the research output constant (i.e. as non-discretionary outputs). We examine how European universities can improve their efficiency in the production of research quality, given the resources they are using and taking into account the level of teaching and research

they produce while moving along a direction which is representative of the median case at European level.

Summing up, we believe that by integrating new data and adopting a novel technique there might be a leap forward in the way in which the activities and performances of universities are examined.

The paper unfolds as follows. Section 2 proposes an outline of the critical literature on university rankings. Section 3 introduces the main sources of data and lists the variables analysed. Section 4 illustrates the methodology and is complemented by Appendix **??** and Appendix A. Finally, Section 5 presents the main results, whilst Section 6 concludes the paper.

2 University rankings. A guided tour of the critical literature

In this Section we organize the main methodological and substantive lines of criticism to university rankings in the four chapters anticipated in Section 1: (a) monodimensionality, (b) statistical robustness, (c) dependence on size and subject mix, and (d) lack of consideration of the input-output structure. Other classifications are certainly possible. For the sake of clarity, criticisms classified in categories (a) and (d) deal with the substantive content of rankings, i.e., the data included (or missing), while studies under (b) and (c) mainly address methodological issues, i.e. how the data are processed in order to arrive at a ranking. Our classification clearly does not exhaust other lines of criticism: for example, we do not have any solution to the issue of English language bias, as well as for the lack of appropriate inclusion of Humanities and Social Sciences in rankings. Also we do not address the more general criticism according to which rankings are a disciplinary device created to impose neoliberal market-oriented values and practices onto an institution, the university, hitherto governed by the public ethos. At the same time our classification is reasonably comprehensive.

2.1 Monodimensionality

The argument is that universities all around the world perform several institutional missions: teaching, research, and third mission. Rankings that programmatically omit the other outputs of universities are therefore biased. Even admitting that the third mission has been legitimized and institutionalized more recently, and is certainly less relevant (quantitatively) than the other two missions, it is felt that ignoring the teaching output altogether severely distorts the reality. Thus there is a demand for including information at least on the teaching output of universities. Existing rankings include only a small set of indicators, whose mean-

ing in terms of overall education activity of universities is questionable: the Alumni Nobel and Field prizes in ARWU (Academic Ranking of World Universities), student/staff ratios (20% weight), international students (5%) and international staff (5%) in QS World University Rankings, and income per academic (2.25%), undergraduates admitted per academic (4.5%), ratio of international to domestic students (2.5%), ratio of international to domestic staff (2.5%) in THE (Times Higher Education Rankings). These proxies are considered unreliable and highy volatile by most analysts, as it is witnessed by the lack of consistency across various rankings, with the exception of the few top positions (Salmi and Saroyan, 2007; Saisana et al., 2011).

In fact, several authors have questioned the correspondence between rankings and quality of education, stating that in general "what is incorporated into the rankings is what is measurable, not what is valid" (Cremonini, Westerhijden and Enders, 2008). The over reliance on research indicators may induce biased decisions (Bastedo and Bowman, 2010).

It is well known that the Shanghai ranking, the first global university ranking, originated from a specific need to provide information on research quality of universities which were considered target for Chinese students and decision makers (Liu, 2009). Therefore it did not incorporate any consideration of the teaching dimension, with the exception of prizes to Alumni, which is however biased towards large and old universities. Other rankings, such as Times Higher Education Supplement, introduced a few items related to education. However, the criticism hits the point: global league tables are largely based on the research output and ignore or underestimate the importance of education (Moed et al., 1985).

Needless to say, including data on education calls into question the issue of quality and in particular on what accounts for quality and whether it can be captured by quantitative measures. Without entering into the theoretical debate, we can say that there is an agreement that data on the completion of studies are an acceptable indicator of quality[1]. While the number of students is certainly an indicator of teaching output (i.e. students are subject to teaching activities during their stay) but not necessarily of quality, the simple fact that a certain proportion of students have achieved a final degree is an acceptable indicator of quality of education.

From a public policy perspective, it would be important to consider that two universities, ranked similarly with respect to research excellence, have largely different social importance depending on the number of students who receive a degree from them, that is, who have

[1]In our case, the information on degrees is the only available quantitative proxy for teaching quality, based on comparable data coming from national statistical authorities at European level. Indeed, comparable data at European level on placement of students would be a better proxy of teaching quality, but unfortunately are not available. Nevertheless, several studies in efficiency analysis suggest to focus on educational degrees. According to Johnes (2006) degrees include elements of quality, since they are the result of the completion of the curriculum. This line of reasoning has also been followed by Daghbashyan et al., (2014).

completed the curriculum. In fact, education is one of the avenues through which new knowledge generates an impact on society.

2.2 Statistical robustness

From the methodological point of view, rankings collapse a variety of indicators into a single measure. This raises a number of technical issues that are the subject of disciplines such as statistics, information theory and decision theory. According to several authors, the validity, reliability and comparability of information incorporated into the measures fail to satisfy properties for acceptance (Bowden, 2000; Van Dyke, 2005; Florian, 2007).

One line of reasoning has stressed the importance of not using just one ranking but multiple ones. More generally, van Leeuven et al. (2003) have underlined the importance of 'using multiple indicators instead of only one' and 'investigating which combinations of indicators provide the strongest correlations, thereby indicating the best combinations of indicators in research performance indicators' (Van Leeuven et al., 2003, p. 276). In a famous and controversial paper, van Raan (2005) warned against the construction of rankings, on the basis of the argument that bibliometric information is biased and subject to errors, so that people do not have 'competence to understand what is measured' (van Raan, 2005, p. 134; see the reply in Liu, Cheng and Lin, 2005).

A second line of research within this chapter has introduced the notion of probabilistic ranking. According to Lubrano (2009) an important methodological problem of rankings is that they assume a deterministic setting, while the underlying indicators are average values from distributions. As Goldstein and Spiegelhalter (1996) puts it, on the contrary, 'the mean has no special status' (Goldstein and Spiegelhalter, 1996, p. 395). In other words, rankings suppress the intrinsic variability of indicators at lower levels of aggregation, giving an impression of stable hierarchies among universities, without explicitly testing for the statistical representativeness of differences. As it has been noted 'an overinterpretation of a set of rankings where there are large uncertainty intervals, can lead both to unfairness and to inefficiency and unwarranted conclusions about changes in ranks' (Goldstein and Spiegelhalter, 1996, p. 405).

A third directions has been pioneered by Saisana et al. (2011), who developed a methodology to test the robustness of rankings. Being based on elementary indicators aggregated into composite indicators, rankings utilize only one of a number of possible combinations of indicators and of aggregation rules. One problem, often raised in the literature, is that the weights used for the aggregation of individual indicators are arbitrary and lack theoretical foundation (see e.g. Provan and Abercromby, 2000). Using a simulation technique, Saisana et al. (2011) show that, in general, rankings are robust in the top positions but less reliable elsewhere, that Shanghai rankings are more robust than Times Higher Education Supple-

ment rankings, and that for a certain numbers of universities the variability induced by changes in the construction of the composite indicator is so large that all existing rankings are meaningless.

2.3 Dependence on university size and subject mix

This line of criticism argues that the rankings are not objective, since they systematically favour old and large universities (Hazelkorn, 2007; 2009). In addition, they favor universities in which scientific, technical and medical discipines (STEM) are dominant. It has been shown, in fact, that controlling for differences in the subject mix may lead to completely different rankings.

With respect to size, the existence of a correlation between the output and the impact of publications has been identified since long time (Hemlin, 1996) Basically, most rankings use absolute numbers of publications and citations as the main element.

The issue of subject mix and the disciplinary composition of universities has also been repeatedly raised in the literature (see Toutkoushian and Webber, 2011 for a discussion). Different disciplines have largely different distributions of scientific output According to Bornmann, de Moya Anegon and Mutz (2013) universities that focus on disciplines such as life sciences have an advantage over universities with a larger importance of disciplines such as engineering, simply because the former have a higher citation volume than the latter. As a consequence, according to several authors (see e.g. Buela-Casal et al., 2007), there should be separate individual rankings for each school or department, rather than having a composite measure. Marginson (2007) has proposed a general principle: 'when comparing research and scholarly capacity or performance, use primarily discipline-based measures rather than whole of institution measures' (Marginson, 2007, p. 19). One important reason to work in this direction is that rankings give a premium to comprehensive research universities. Isomorphic pressures may reduce the diversity of the system penalizing programmatic diversity and specialist universities (Marginson and van der Wende, 2007). Thus the issue here is not to use several rankings or to check their robustness or to avoid aggregation but rather use separate disciplinary rankings.

2.4 Lack of consideration of the input-output structure

Another line of criticism argues that rankings simply ignore the amount of resources that universities receive. According to OECD data, governments allocate to higher education widely different amount of resources, resulting in large gaps in student/staff ratios, as well as in cost per student (Porter and Toutkoushian, 2006). Accordingly, it is argued that rankings are, at least partially, a reflection of the economic status of countries. If this is

the case, they would give no information as to how to improve the system within countries (Docampo, 2012). Furthermore, they might lead to wrong implications for the allocation of resources (Stake, 2006).

Bornmann, Lutz and Daniel (2013) have shown that 80% of the variance between the universities is explained by differences between the countries in which the universities are located, in particular by differences in GDP per capita. This leads to asking whether rankings measure the differential performance of universities, or rather reflect the divide in scientific performance among countries, a factor upon which individual universities have little power. A related and subtle criticism has been proposed by Cremonini, Westerheijden and Enders (2008), who argue that rankings want to reframe higher education as a consumer good, while the appropriate reference model should be one of investment. In other words, rankings offer only information on the output, while they fail to account for the relation between inputs and outputs, and between outputs and social outcomes.

Safon (2013) has shown that the position in rankings is largely determined by underlying factors such as "age, scope, activity in hard sciences, university in U.S., English-speaking country, annual income, orientation towards research, and reputation" (p. 238). As it is clear from this list, only a few of these factors, such as orientation towards research and, partially, reputation, are under the control of universities, while others mostly depend on historical factors (age, scope and activity in hard sciences) or on country-level factors (English and annual income). While it will not be possible to control for all contextual factors and isolate those that are under the control of universities (an issue that has been prominent for decades in the literature in industrial economics and strategic management and is still largely unsolved), some improvement can be pursued.

As a matter of fact, most of these authors challenge the notion that rankings can be built mainly on the basis of output data. Rather, the appropriate notion to be used in order to compare universities is the one of efficiency, or the relation between input and output. The joint consideration of outputs and resources employed is the starting point for university strategy (Bonaccorsi and Daraio, 2007) and for the positioning of universities with respect to their peers (Bonaccorsi and Daraio, 2008).

3 Data

We exploit a large database, recently constructed by the European University Micro Data (EUMIDA) Consortium under a European Commission tender, supported by DG EAC, DG RTD, and Eurostat.

This database is based on official statistics produced by National Statistical Authorities in all 27 EU countries (with the exception of France and Denmark) plus Norway and

Switzerland. The EUMIDA project, relying on the results of the Aquameth project (Bonaccorsi and Daraio, 2007; Daraio et al. 2011) included two data collections: Data Collection 1 (DC 1) included all higher education institutions that are active in graduate and post-graduate education (i.e. universities), but also in vocational training. Data refer to 2008, or to 2009 in some cases. Thus all institutions delivering ISCED (International Standard Classification of Education) 5a and 6 degrees are included, and the subset of those delivering ISCED 5b degrees that have a stable organization (i.e. mission, budget, staff). There are 2457 institutions identified in Data Collection 1: these constitute the perimeter of higher education institutions in Europe. On these institutions a large set of uniform variables have been collected.

Of these, 1364 are defined research active institutions: of these only 850 are also doctorate awarding. They are the object of Data Collection 2 (DC 2), for which a larger set of variables were collected. This means that a significant portion of research active institutions is found outside the traditional perimeter of universities, that is in the domain of non-university research (particularly in countries with dual higher education systems).

We integrate the EUMIDA data, in particular the DC 2 dataset, with the Scimago data (SIR World Report 2011, period analyzed 2005-09) that include institutions having published at least 100 scientific documents of any type, that is, articles, reviews, short reviews, letters, conference papers, etc., during the period 2005-2009 as collected by Scopus database. From Scimago data we used the following variables:

- number of publications in Scopus (PUB);

- Specialization index (SPEC) of the university that indicates the extent of thematic concentration / dispersion of an institution's scientific output; its values range between 0 to 1, indicating generalistic vs. specialized institutions respectively. This indicator is computed according to the Gini Index and in our analysis it is used as a proxy of the specialization of the university.

- International Collaboration (IC), % of a university's output realized in collaboration with foreign institutions (calculation based on affiliations with more than one country address).

- High Quality Publications (Q1), % of publications that a university publishes in the first quartile (25%) in their categories as ordered by Scimago Journal Rank indicator.

- Normalized Impact (NI), in % shows the relationship between an university's average scientific impact and the world average set to a score of 1.

- Excellence Rate (EXC), % of university output that is included in the 10% of the most cited paper in their respective scientific fields.

Table 1 defines and describes the inputs, outputs and conditioning factors that are used in the following analysis.

Table 1: Definition of inputs, outputs and conditioning factors

Input/Output/Conditioning factor	**Definition**
Input	
NACSTA	Number of non academic staff
ACSTAF	Number of academic staff
PEREXP	Personnel expenditures (PPS)
NOPEXP	Non-personnel expenditures (PPS)
FINP	Input factor including:
	NACSTA, ACSTAF, PEREXP, NOPEXP
Output	
TODEG5	Total Degrees ISCED 5
TODEG6	Total Degrees ISCED 6 (Doctorate)
PUB	Number of published papers (Scimago)
IC	International collaboration (Scimago)
NI	Normalized impact (Scimago)
Q1	High quality publications (Scimago)
EXC	Excellence rate (Scimago)
FRES	Factor of research including: TODEG6, PUB
FQUAL	Factor of quality of research including:
	IC, NI, Q1, EXC
Conditioning factors	
SIZE	It is the log of the sum of
	Total Students enrolled at both ISCED
	5 and ISCED 6 level
SPEC	Proxy of Specialization
	Gini index of the scientific output (Scimago)

Source: Eumida DC2 and Scimago.

As usually used in applied econometrics, the size is computed as the logarithm of the total volume of the activity, that in our case is proxied by the sum of enrolled students at all undergraduate and post-graduate levels.

4 Directional Distances, Conditional Distances and Managerial efficiencies

4.1 Basic concepts and notations

We model European universities in a production activity framework. In this setup, universities are the producing units (hereafter 'units') and produce a set of outputs $Y \in \mathbb{R}^q$ by combining a set of resources (inputs) $X \in \mathbb{R}^p$. The production activity is characterized by the attainable set Ψ, the set of combination of the production plans (x, y) that are technically achievable:

$$\Psi = \{(x, y) \in \mathbb{R}^p \times \mathbb{R}^q | x \text{ can produce } y\}. \tag{4.1}$$

We know (see Daraio and Simar, 2007) that the set Ψ can be described as:

$$\Psi = \{(x, y) \in \mathbb{R}^p \times \mathbb{R}^q | H_{XY}(x, y) > 0\}, \tag{4.2}$$

where $H_{XY}(x, y)$ is the probability of observing a unit (X, Y) dominating the production plan (x, y), i.e. $H_{XY}(x, y) = \text{Prob}(X \leq x, Y \geq y)$.

The efficient boundary of Ψ is of interest and several ways have been proposed in the literature to measure the distance of the unit (x, y) to the efficient frontier. One of the most flexible approach is the directional distance introduced by Chambers et al. (1996). Given a directional vector for the inputs $d_x \in \mathbb{R}^p_+$ and a direction for the outputs $d_y \in \mathbb{R}^q_+$, the directional distance is defined as:

$$\beta(x, y; d_x, d_y) = \sup\{\beta > 0 | (x - \beta d_x, y + \beta d_y) \in \Psi\}, \tag{4.3}$$

or equivalently (as reported also in Daraio and Simar, 2014[2]):

$$\beta(x, y; d_x, d_y) = \sup\{\beta > 0 | H_{XY}(x - \beta d_x, y + \beta d_y) > 0\}. \tag{4.4}$$

Hence, we measure the distance of unit (x, y) to the efficient frontier in an additive way and along the path defined by $(-d_x, d_y)$.

This way of measuring the distance generalizes the 'oriented' radial measures proposed by Farrell (1957). Indeed by choosing $d_x = 0$ and $d_y = y$ (or $d_x = x$ and $d_y = 0$), we recover the traditional output (respectively input) oriented radial distances. The flexibility of this approach relies in the possibility of setting some elements of the vector d_x and/or of the vector d_y to zero, for focusing on the distances to the frontier along certain particular paths (for instance if some inputs or outputs are non-discretionary, not under the control of the units, etc.).

[2]See the references cited there.

Consistent nonparametric estimators of Equation 4.4 can be found in Daraio and Simar (2014) which analyzes in details the case when some directions are set to zero, as well as statistical issues in this context.

4.2 Modeling strategy

1. Multi-input multi-output activity of universities
   The approach described in Section 4.1 permits to model the activity of universities as multi-input multi-output production units. Ideally, we would like to compare European universities taking into account all their outputs of teaching, research and 'third mission'. The data described in Section 3 are of great value at this purpose; however in our database third mission dimensions have a low coverage and for that reason where excluded. We run an exploratory data analysis[3] and given the high correlations observed among variables (all greater than 85%) we ended up with the following variables to proxy the activity of universities. One input, FINP (a factor including NACSTA, ACSTAF, PEREXP, NOPEXP); and three outputs: TODEG5 (proxy of the teaching activity), FRES (a factor of research including PUB and TODEG6) and FQUAL (a factor of quality of research including IC, NI, Q1 and EXC). See Table 1.

2. Target setting
   For a discussion about the choice of a direction to approach the efficient frontier, see Färe et al. (2008). The direction can be different for each unit (like in the radial cases) or it can be the same for all the units. Färe et al. (2008) argue that a common direction would be a kind of egalitarian evaluation reflecting some social welfare function.

   In this paper we select the same direction for all the units, setting a reference with respect to the European standard. The reference is made with respect to the median value calculated at European level on the analysed sample.

   We adopt then an *output directional distance* in which the inputs are given (FINP), two outputs TODEG5 and FRES are non-discretionary (that means that are considered in the estimation of the production possibility set Ψ but are not active in the maximization) and one output, FQUAL, is the target. This means that universities are compared on their ability to produce quality of research (FQUAL, a factor of IC, NI, Q1 and EXC), given the inputs used and taking into account their teaching activity (TODEG5) and the volume of their research (FRES).

   In this paper we attempt to investigate how European universities are doing in the production of Excellent science, a pillar of the European Research Area. This attempt

[3]The details are not reported to save space.

is possible thanks to the availability of comparable micro-data on European universities and their integration with the scientific production outputs described in Section 3. The path along which we compare European university performance to reach the efficient frontier is the same for all universities and corresponds to the median value of FQUAL, computed at European level.

See Figure 1 for an illustration. In Figure 1 stars are the units and the arrows show the path of units to reach the efficient frontier; u is a university and u' its projection onto the efficient frontier: given its value of TEACH and FRES (non-discretionary outputs), the unit has to improve in the production of FQUAL going from u towards u'.

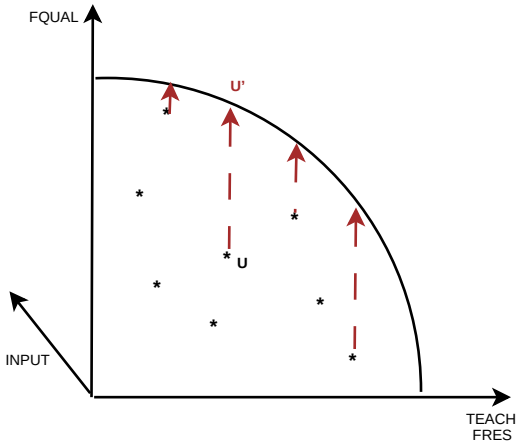

Figure 1: *The production model of excellent science by European universities: a simplified illustration.*

It may be useful for policy makers to measure, in original units of the outputs, the estimated distance of a unit to the frontier. This allows to appreciate the efforts to be achieved in increasing the outputs and decreasing the inputs to reach the efficient frontier. This measure is given by what we call the 'gaps' to efficiency. They are directly given by:

$$G_x = \widehat{\beta}(x, y; d_x, d_y)d_x, \quad \text{and} \quad G_y = \widehat{\beta}_\alpha(x, y; d_x, d_y)d_y. \tag{4.5}$$

3. Taking size and subject mix into account

From the literature review reported in Section 2 we know that both size and scientific specialization (SPEC) have a significant impact on the performance of European universities. In our model then we will condition the efficiency estimation to these factors,

to account for their influence on the distribution of inefficiency, that is the distance of the units from the efficient boundary. In Section 4.4 we detail how to include these factors in the directional distance framework described above.

We aim at comparing how European universities are doing in the production of Quality of Research (FQUAL), given their inputs and taking into account their teaching (TODEG5) and their volume of research (FRES), the latter outputs considered as non discretionary outputs, and conditioning the comparison to the impact of SIZE and SPEC.

4. A 'FAIR' COMPARISON OF UNIVERSITY PERFORMANCE The conditional directional methodology described in Section 4.4 is useful also to make a step further in the comparison of European university performance. As illustrated in Section 4.5, we can 'depurate' the efficiency scores from the influence of SIZE and SPEC to compare Universities in Europe on the base of their 'managerial' ability, that is measured as the *residual* of the conditional efficiency score net of SIZE and SPEC effects.

5. ACCOUNTING FOR STATISTICAL ROBUSTNESS One of the main methodological issue of the literature on rankings is the statistical robustness of the proposed approach. In this paper we account for statistical robustness by applying directional distances estimators robust to extremes and outliers (see Section 4.3) and estimating bootstrap error bounds on the managerial efficiency scores (see Appendix A).

4.3 Robust Directional Distances

Quantile frontiers for evaluating the performance of units by using oriented radial measures (input or output) are currently used (see Simar and Wilson 2014 for a recent survey). Their adaptation to directional distance is quite natural after the representation given in (4.4). In place of looking to the support of the distribution H_{XY} we benchmark the unit against a point which leaves on average $\alpha \times 100\%$ of points above the frontier. This benchmark is the α-quantile frontier. Formally the α-order directional distance is defined as

$$\beta_\alpha(x, y; d_x, d_y) = \sup\{\beta > 0 | H_{XY}(x - \beta d_x, y + \beta d_y) > 1 - \alpha\}. \tag{4.6}$$

Here a value $\beta_\alpha(x, y; d_x, d_y) = 0$ indicates a point (x, y) on the α-quantile frontier, a positive value is a point below the quantile frontier and a negative value is a point above the quantile frontier. We see clearly that when $\alpha \to 1$ we recover the full frontier definition.

A nonparametric estimator[4] can be founded in Daraio and Simar (2014) which detail the case with some non-discretionary inputs and/or outputs that we apply in this paper.

[4]Called hereafter $\beta_{\alpha,FDH}$ because it is the robust version of a directional distance based on a nonconvex FDH (Free Disposal Hull, Deprins et al. 1984) estimator.

The projection of any $(x, y) \in \Psi$ on the estimated α-quantile frontier is given by the points $(\hat{x}_\alpha^\partial, \hat{y}_\alpha^\partial)$ defined as

$$\hat{x}_\alpha^\partial = x - \widehat{\beta}_\alpha(x, y; d_x, d_y)d_x, \quad \text{and} \quad \hat{y}_\alpha^\partial = y + \widehat{\beta}_\alpha(x, y; d_x, d_y)d_y. \tag{4.7}$$

Since the resulting estimator will not envelop all the data points, the resulting frontier is more robust to outliers and extreme data points than its full version above.

For the partial frontiers, the gaps appear as being the difference between (x, y) and the projections on the α-quantile frontier given in (4.7). They are particularly useful to detect outliers in the direction given by (d_x, d_y). This will be the case in the input direction if $G_{\alpha,x} = \widehat{\beta}_\alpha(x, y; d_x, d_y)d_x$ has some elements with large negative value: the point (x, y) is well below the estimated α-frontier in the input direction, and/or a very large negative value in some elements of the vector $G_{\alpha,y} = \widehat{\beta}_\alpha(x, y; d_x, d_y)d_y$ warns a point being well above the quantile frontier.

It is well known that nonparametric efficiency analysis gain in precision when working in space with lower dimensions (this is the usual "curse of dimensionality" of nonparametric techniques, see e.g. Daraio and Simar (2007), for a discussion). In our application, the original data are transformed before entering into the analysis, to reduce the dimension of the problem (by using input and/or output factors as defined in Daraio and Simar, 2007, p. 148 and followings). In this case, once the gaps have been computed for the variables used in the analysis, there is a need to evaluate the corresponding gaps in the original inputs and outputs. This can be achieved by transforming back the gaps in the factors into the original units. For more details, see the Appendix of Bonaccorsi, Daraio and Simar (2014).

4.4 Conditional directional distances

In this section we introduce in the production model described above external or environmental factors $Z \in \mathbb{R}^r$. These variables are neither inputs nor outputs, and they are not under the direct control of the manager. However, they may influence the production process. A natural way for introducing these variables through conditional efficiency measures could be as follows[5].

The idea is very simple, we only have to replace $H_{XY}(x, y)$ in the above unconditional model by $H_{XY|Z}(x, y|Z = z) = \text{Prob}(X \leq x, Y \geq y|Z = z)$ where we condition to the value z of the external factors that the unit (x, y) has to face. In our setup here, this permits to define a conditional directional distance $\beta(x, y; d_x, d_y|z)$. Daraio and Simar (2014) provide a non parametric estimator of $H_{XY|Z}(x, y|Z = z)$ when some directions are set to zero as well as its robust version[6] that we apply in this paper.

[5] See Daraio and Simar (2007) for more details.

[6] Called hereafter $\beta_{\alpha,FDH|Z}$.

This approach has been applied to our European university data for including SIZE and SPEC in the multidimensional evaluation of university performance.

4.5 Estimation of Managerial efficiency scores

Many of the existing studies for investigating the effect of external environmental factors are based on simple two-stage regression analyses where estimated efficiency scores (input or output oriented) are regressed in a second stage against the Z variables. However we know from the literature (Simar and Wilson, 2007) that this is valid only under a restrictive 'separability' assumptions where it is assumed that the frontier of the attainable set is not changing with the values of z. As indicated in Badin, Daraio and Simar (2012), the use of the estimated conditional efficiency scores for this second stage regression, does not requires this assumption. We can evidently do the same here with conditional directional distances. The flexible second stage regression can be written as the following location-scale nonparametric regression model (the presentation here follows Daraio and Simar, 2014):

$$\beta(X, Y; d_x, d_y | Z = z) = \mu(z) + \sigma(z)\varepsilon, \tag{4.8}$$

where $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = 1$ and

$$\mu(z) = \mathbb{E}(\beta(X, Y; d_x, d_y | Z = z)) \text{ and } \sigma^2(z) = \mathbb{V}(\beta(X, Y; d_x, d_y | Z = z)).$$

These two functions can be estimated non parametrically from a sample of observations $\left\{Z_i, \widehat{\beta}(X_i, Y_i; d_x, d_y | Z_i)\right\}$, $i = 1, \ldots, n$ by using, e.g., Nadaraya-Watson or local linear estimates (see Daraio and Simar, 2014 for technical details). As shown with simulated samples in Badin et al. (2012), the analysis of $\widehat{\mu}(z)$ as a function of z will enlighten the potential effect of Z on the average efficiency, with the help of $\widehat{\sigma}(z)$ which may indicate the presence of heteroskedasticity.

An important result of the above approach is the analysis of the 'residuals'. For a given unit we can define the error term:

$$\varepsilon = \frac{\beta(X, Y; d_x, d_y | Z = z) - \mu(Z)}{\sigma(Z)} \tag{4.9}$$

This can be viewed as the *unexplained* part of the conditional efficiency score. If Z is independent of ε, this quantity can be interpreted as a 'pure' or 'managerial' efficiency measure of the unit since it is the remaining part of the conditional efficiency after removing the location and scale effect due to Z. It is called 'managerial' because it depends only upon the managers of units ability and not upon the environmental factors, and it represents an advanced and robust interpretation of the Leibenstein's (1966) X-inefficiency theory.

The label 'managerial' does not convey any analogy between universities and private firms. We fully recognize that universities are not maximizing an objective function under conditions of competition. We also recognize that universities are, at least in most European countries, governed by a complex governance in which the academic side is dominant with respect to management and/or stakeholders. Thus the label 'managerial' should not be interpreted literally. The label only emphasizes that part of the efficiency of universities may depend on internal decisions, with respect to adjustments in inputs (e.g. recruitment of academic staff) or outputs (e.g. offering of new courses in specific disciplines).

What we have done is a kind of *whitening* of the conditional efficiency scores, from the effects due to the environmental-external conditions Z. We can use these quantities (the estimated ε, indicated as $\widehat{\varepsilon}$), which are standardized (mean zero and variance one), to compare the units among them on a fair base: a large value of $\widehat{\varepsilon}$ indicates a unit which has poor performance, even after eliminating the main effects of the environmental factors. A small (negative) value, on the contrary, indicates very good managerial performance of the unit. It allows to rank the units facing different external conditions (SIZE and SPEC), because the main effects of these factors have been eliminated. Extreme (unexpected) values of $\widehat{\varepsilon}$ would also warn for potential outliers. Of course, the above analysis could also be performed by using robust efficiency scores; for instance the selection of a value of α near 1 provides a robust version of the above analysis, and this was done in the empirical illustration that follows.

Finally, to provide an investigation on the sampling variations (due to the nonparametric model applied in this paper), we apply the bootstrap for estimating error bounds on the managerial efficiency scores. For more details, see the Appendix A.

5 Results

In this section we summarize the main results obtained from our analysis on the European university dataset described in Section 3.

Table 2 reports in the columns: Country, number of observations (# obs), number of dominating units (# dom), empirical estimates of the probability of being dominated ($\widehat{H}_{XY}$), robust directional measure of efficiency ($\beta_{\alpha,FDH}$) and robust directional measure of efficiency conditioned to SIZE and SPEC, our Z variables ($\beta_{\alpha,FDH|Z}$).

The last line of the Table shows the average at European level. An outline of the efficiency analysis results could be obtained by comparing the average performance at national level with the European average. We recall that the values of the efficiency scores have to be interpreted as follows: the lower their values the higher the performance is. Countries that are performing much better than the European standard are UK, Sweeden and Switzerland,

followed by Belgium, Austria, Ireland and Netherlands, this appears if we consider both the unconditional efficiency scores ($\beta_{\alpha,FDH}$) and the conditional efficiency scores ($\beta_{\alpha,FDH|Z}$).

Table 2: Efficiency Results: averages by country.

Country	#obs	#dom	$\widehat{H}_{XY}$	$\beta_{\alpha,FDH}$	$\beta_{\alpha,FDH\|Z}$
AT	13	4.46	0.0111	0.049757	0.080548
BE	3	3.33	0.0083	0.028793	0.083364
CH	10	1.20	0.0030	-0.105144	0.009617
CZ	11	3.55	0.0088	0.201562	0.196404
DE	62	11.94	0.0298	0.225836	0.239438
ES	42	6.76	0.0169	0.181811	0.172416
FI	5	2.80	0.0070	0.115628	0.147933
HU	6	27.50	0.0686	0.354777	0.371831
IE	6	3.33	0.0083	0.023005	0.104317
IT	45	5.31	0.0132	0.096320	0.142867
NL	7	5.57	0.0139	0.085189	0.130575
NO	8	6.25	0.0156	0.140668	0.145187
RO	7	2.71	0.0068	0.171315	0.255163
SE	9	2.33	0.0058	-0.077888	0.055961
UK	73	1.97	0.0049	-0.067431	0.063181
EU	313	6.07	0.0151	0.092477	0.145114

Note: only countries with at least 3 observations are reported in the table.
The last line reports the average over the whole analyzed sample.

Table 3 reports the estimated gaps in percentage of the outputs produced by the units. As expected, countries that perform better on average are again: UK, Sweeden and Switzerland, followed by Belgium, Austria, Ireland and Netherlands. Now also Norway perform slightly higher than the European average.

Table 3: Gaps in percentages: averages by country.

Country	#obs	#DEG5	#DEG6	#PUB	IC	$Q1$	NI	EXC
AT	13	0.00	0.00	0.00	0.06	0.08	0.07	0.10
BE	3	0.00	0.00	0.00	0.07	0.08	0.08	0.07
CH	10	0.00	0.00	0.00	0.01	0.01	0.01	0.01
CZ	11	0.00	0.00	0.00	0.26	0.39	0.33	0.64
DE	62	0.00	0.00	0.00	0.25	0.25	0.24	0.27
ES	42	0.00	0.00	0.00	0.21	0.19	0.21	0.25
FI	5	0.00	0.00	0.00	0.17	0.21	0.18	0.28
HU	6	0.00	0.00	0.00	0.38	0.41	0.57	0.49
IE	6	0.00	0.00	0.00	0.09	0.15	0.12	0.18
IT	45	0.00	0.00	0.00	0.19	0.14	0.16	0.18
NL	7	0.00	0.00	0.00	0.11	0.13	0.11	0.15
NO	8	0.00	0.00	0.00	0.14	0.16	0.14	0.21
RO	7	0.00	0.00	0.00	0.37	1.17	0.57	2.35
SE	9	0.00	0.00	0.00	0.05	0.06	0.06	0.07
UK	73	0.00	0.00	0.00	0.08	0.07	0.07	0.09
EU	313	0.00	0.00	0.00	0.17	0.19	0.17	0.26

Note: only countries with at least 3 observations are reported in the table.
The last line reports the average over the whole analyzed sample.

Figure 2 illustrates the distribution of the Managerial efficiency scores estimated over the whole European sample. We remind that large values indicate units which have poor performance even after eliminating the main effects of SIZE and SPEC. Small or negative values indicate instead good managerial performance of the units. It is interesting to note that the global distribution of the European managerial efficiency estimated over our sample shows a peak around -0.9, a value that indicates very good managerial performance.

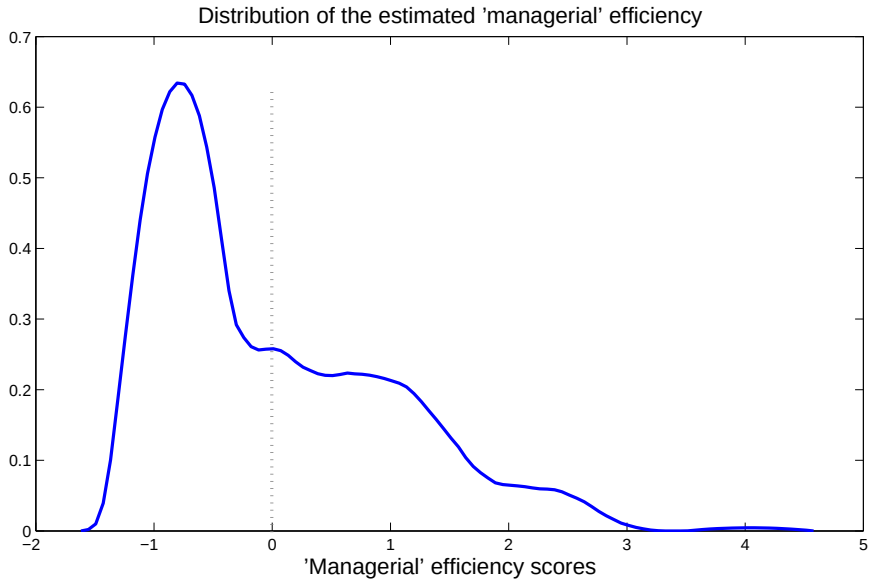

Figure 2: *Nonparametric kernel distribution of the estimated Managerial Efficiency scores.*

The following graphics, from Figure 3 to Figure 17 report the estimated managerial efficiency scores by country and show that a certain variability within countries exist even if a few countries seem to have a superior managerial efficiency, namely, Belgium, Netherlands and Switzerland.

By inspecting the figures it is clear that in each country there are universities below and above the horizontal line. It would be interesting, in future studies, to examine these cases individually and to identify ways for improving the condition of inefficient universities vis-a'-vis their national peers and within the respective national context. These universities are strictly comparable in a multi-input multi-output framework, because their input conditions (for example the expenditure for personnel) are equalized at national level. At the same time the figures also show that there are countries in which only a small share of universities are inefficient. In other words, there are country-level factors that enhance the capabilities of universities to adapt to external conditions and to make the best use of their resources.

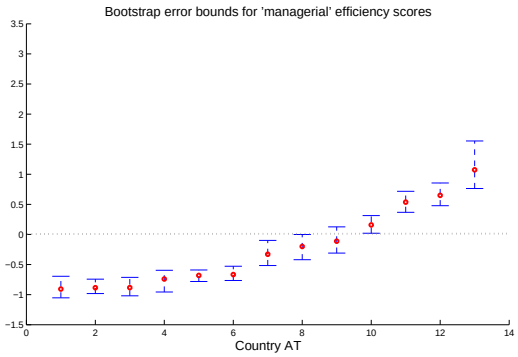

Figure 3: *AT Managerial efficiency scores with bootstrap error bounds.*

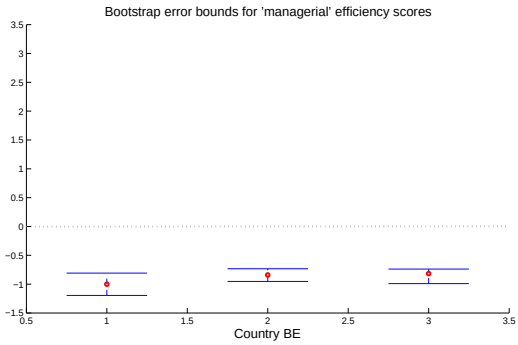

Figure 4: *BE Managerial efficiency scores with bootstrap error bounds..*

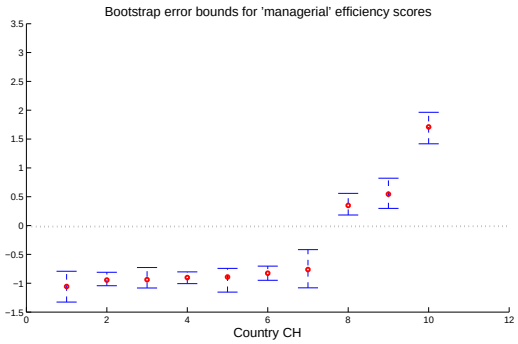

Figure 5: *CH Managerial efficiency scores with bootstrap error bounds.*

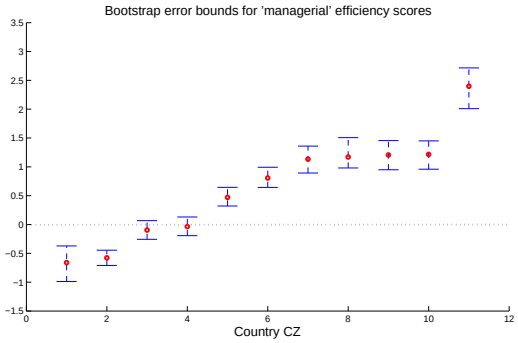

Figure 6: *CZ Managerial efficiency scores with bootstrap error bounds.*

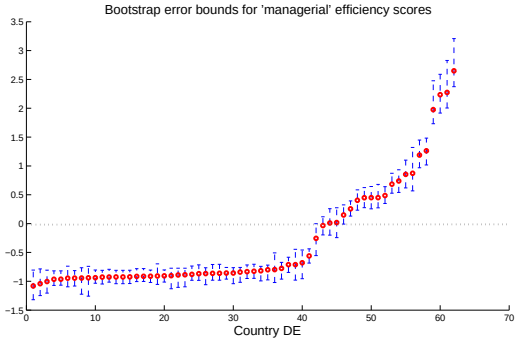

Figure 7: *DE Managerial efficiency scores with bootstrap error bounds.*

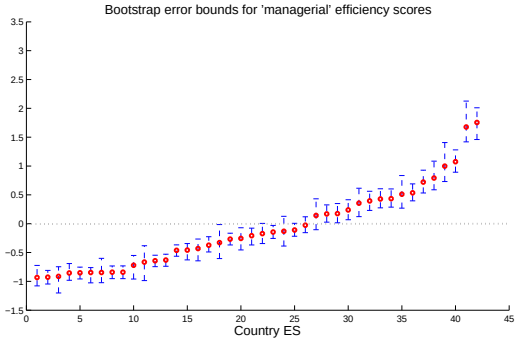

Figure 8: *ES Managerial efficiency scores with bootstrap error bounds.*

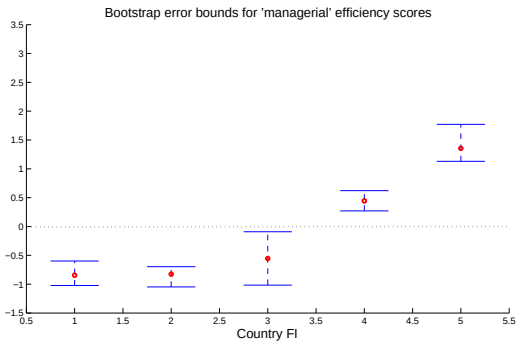

Figure 9: *FI Managerial efficiency scores with bootstrap error bounds.*

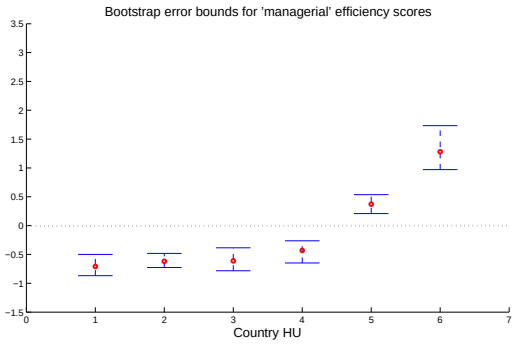

Figure 10: *HU Managerial efficiency scores with bootstrap error bounds.*

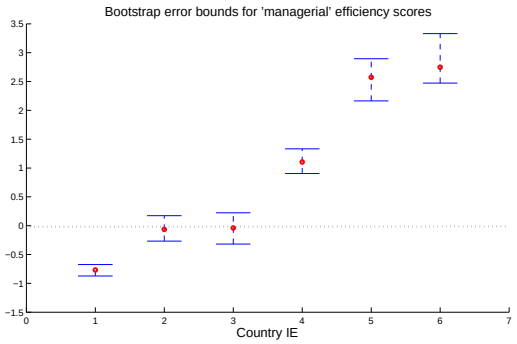

Figure 11: *IE Managerial efficiency scores with bootstrap error bounds.*

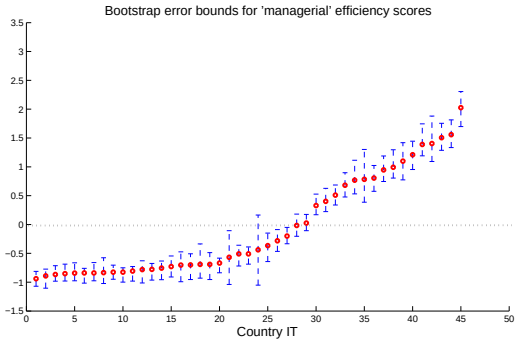

Figure 12: *IT Managerial efficiency scores with bootstrap error bounds.*

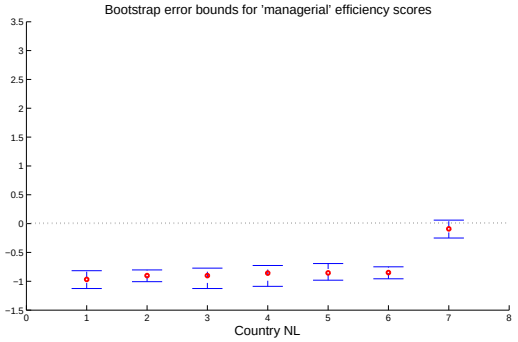

Figure 13: *NL Managerial efficiency scores with bootstrap error bounds.*

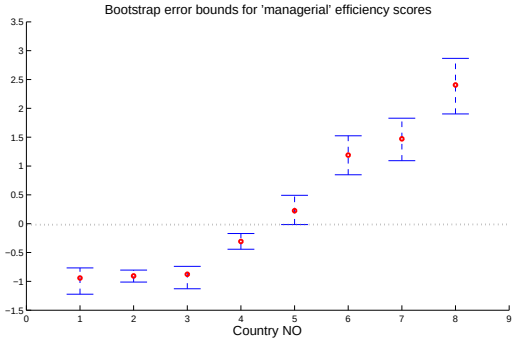

Figure 14: *NO Managerial efficiency scores with bootstrap error bounds.*

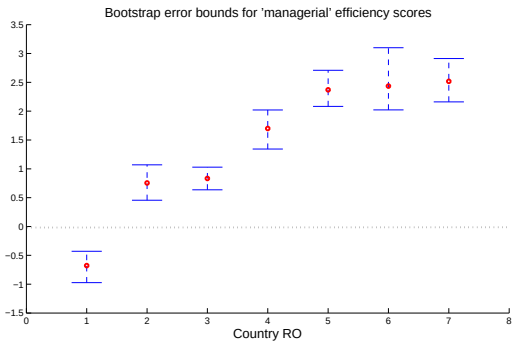

Figure 15: *RO Managerial efficiency scores with bootstrap error bounds.*

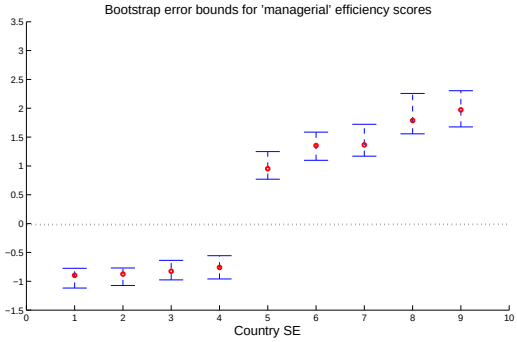

Figure 16: *SE Managerial efficiency scores with bootstrap error bounds.*

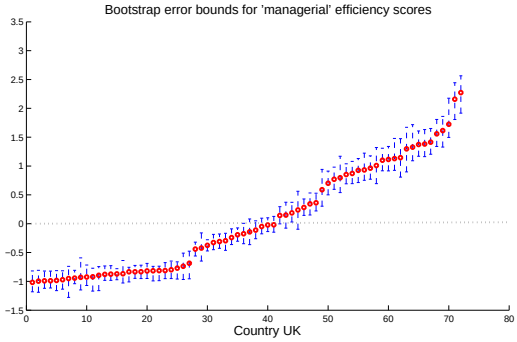

Figure 17: *UK Managerial efficiency scores with bootstrap error bounds.*

6 Conclusions

We provide a first attempt to overcome current limitations of existing rankings by using an original and comprehensive database on European universities microdata integrated with bibliometric data on the scientific production, and by applying recently developed techniques in efficiency analysis.

Our results are encouraging and clearly show that investing in data integration and opening of the available data to the research community and to policy makers would certainly improve our current state of the art methodologies and empirical evidences on European universities position in their multidimensional space of activities and performance.

The estimation of the managerial efficiency scores has shown a large variability within countries. This means that there is room for many universities to improve their performance. More precisely, these universities, keeping the national institutional framework and legislation constant, might increase the production of high quality research, without compromising their volume of research and the quality of education, as proxied by the number of degrees they deliver. This is possible simply because other universities, in the same national context, perform better. This is an important result, which should be contrasted with the state of the art of rankings. Existing university rankings are monodimensional, being based largely of research output. This leaves unanswered the question as to the trade-offs that universities face between improving the quality of research and delivering education. Our data carefully disentangle different dimensions of performance (education, volume of research, quality of research) and allow the identification of a clear direction for improvement.

At the same time, the inspection of average efficiency values per country shows large differences due to the national context. The interpretation of these differences will require a dedicated effort. In particular, additional efforts should be directed to comparability, validation and data quality of European data on higher education institutions as well as to the opening and integration of these data in broader platforms.

Nevertheless, a preliminary conjecture could be as follows. In order to make the best use of their inputs, universities should be put in the position to move in the multidimensional strategic space. This space includes inputs and outputs. Efficient universities are those that adjust their mix of inputs in order to achieve the best possible mix of outputs. It is clear that universities do not have full discretionary power over inputs and outputs, as our analysis has clearly recognised. However, national contexts may provide more or less strategic autonomy, that is, may support universities in their strategic positioning or may, on the contrary, create legal and administrative constraints.

Supporting the autonomy of universities in strategic positioning is generally associated to two conditions. As for education, it requires that universities are in the position to match appropriately the profile of students to the teaching offering. While this may have

different implications in different fields, there is a well known general problem that cuts across fields of education and countries, namely the role of professional education, also called vocational training. Some countries allocate vocational training to separate institutions, while others add to the general mission of universities. In the latter case universities have, in general, larger student loads and lower teaching efficiency, given the mismatch between the educational needs of students and the rigidity of the university offering.

As for research, efficiency requires that government research funding is allocated according to criteria that gives a premium to research quality. This follows the adoption of evaluation exercises, or formula-based funding criteria based on research quality. Universities that are placed in an institutional context based on research quality funding develop over time strategies to improve their positioning in research.

These two conditions can also be described as differentiation, respectively in education and in research. National systems differentiated in education include dual and binary systems, as adopted in countries of German tradition and in Scandinavian countries. National systems differentiated in research include countries, such as United Kingdom, Netherlands and Switzerland, and more recently other Scandinavian countries, in which there is not legal segregation among university institutions (as it happens in France), but *de facto* vertical differentiation along the research dimension, based on differential access to research funding.

We therefore advance the conjecture that countries with a higher efficiency of universities, net of size and subject mix, are those that are more differentiated. Netherlands and Switzerland are countries with differentiation in both education and research; United Kingdom is highly differentiated in research (while vocational training is carried out only by poorly performing universities in research, or de facto delegated to the private sector, creating an effect of differentiation without legal segregation); Sweden is differentiated in education and has moved more recently but aggressively towards differentiation in research.

If this conjecture would be confirmed, it would be consistent with other studies based on the EUMIDA dataset (Bonaccorsi, 2014) and the previous Aquameth dataset (Daraio et al., 2011; Bonaccorsi and Daraio, 2007).

A Appendix: Bootstrap Error Bounds for Managerial Efficiency scores

We use standard bootstrap methods (for an introduction, see Efron and Tibshirani, 1993) for building prediction intervals for the pure efficiencies. The bounds of these prediction intervals are also called the bootstrap error bounds. The 'managerial' efficiencies are estimated as the residual of the nonparametric location-scale regression of the efficiency scores $\widehat{\beta}_\alpha(X_i, Y_i|Z_i)$

on the variables Z_i

$$\widehat{\beta}_\alpha(X_i, Y_i|Z_i) = \mu(Z_i) + \sigma(Z_i)\varepsilon_i,\ i = 1, \ldots, n, \tag{B.1}$$

where $\mathbb{E}(\varepsilon_i|Z_i) = 0$ and $\mathbb{V}(\varepsilon_i|Z_i) = 1$. In Daraio and Simar (2014), it is shown why the bootstrap can be used for inference in the nonparametric regression of $\widehat{\beta}_\alpha(x, y|z)$ on z. This is typically due to the fact that the order-α estimators do not suffer from the curse of dimensionality attached to x and y. The same argument obviously applies here.

So, the nonparametric estimation of the model (B.1) produce $\widehat{\mu}(z)$ and $\widehat{\sigma}(z)$ for any z and the resulting residuals

$$\widehat{\varepsilon}_i = \frac{\widehat{\beta}_\alpha(X_i, Y_i|Z_i) - \widehat{\mu}(Z_i)}{\widehat{\sigma}(Z_i)} \tag{B.2}$$

are interpreted as the 'managerial efficiency measures' for the units $i = 1, \ldots, n$.

This is a pointwise predictor of the random variable ε_i and we would like to build a confidence interval (more precisely a prediction interval) of a given level (say, 95%) for each unit. We adapt here in the nonparametric model (B.1), the procedure described in Simar and Wilson (2010) for parametric models. The algorithm can be summarized as follows:

[1] Rescale the residuals to obtain residuals with mean zero and variance 1:

$$\widetilde{\varepsilon}_i = \frac{\widehat{\varepsilon}_i - \overline{\widehat{\varepsilon}}}{\sqrt{n^{-1}\sum_{j=1}^n \left[\widehat{\varepsilon}_i - \overline{\widehat{\varepsilon}}\right]^2}}, \tag{B.3}$$

where $\overline{\widehat{\varepsilon}}$ is the sample mean of the n original $\widetilde{\varepsilon}_i$.

[2] Redo the next steps a large number of B times (e.g. $B = 2000$ is enough for most of the empirical applications) , so $b = 1, \ldots, B$.

[2.1] Draw randomly with replacement n values $\varepsilon_i^{*,b}$ among the n rescaled values $\widetilde{\varepsilon}_i$.

[2.2] For the same values of Z_i generate n bootstrap values of $\beta_\alpha^{*,b}$ as follows

$$\beta_\alpha^{*,b}(X_i, Y_i|Z_i) = \widehat{\mu}(Z_i) + \widehat{\sigma}(Z_i)\varepsilon_i^{*,b},\ i = 1, \ldots, n$$

[2.3] From the bootstrap sample of size n of pairs $(\beta_\alpha^{*,b}(X_i, Y_i|Z_i), Z_i)$ estimate the bootstrap analog of (B.1). We obtain $\widehat{\mu}^{*,b}(Z_i)$ and $\widehat{\sigma}^{*,b}(Z_i)$, for $i = 1, \ldots, n$.

[2.4] Build now the n bootstrap versions of the pure efficiencies as

$$\widehat{\varepsilon}_i^{*,b} = \frac{\widehat{\beta}_\alpha(X_i, Y_i|Z_i) - \widehat{\mu}^{*,b}(Z_i)}{\widehat{\sigma}^{*,b}(Z_i)},\ i = 1, \ldots, n. \tag{B.4}$$

[3] At the end of step [2] we have B bootstrap values $\widehat{\varepsilon}_i^{*,b}$, $b = 1, \ldots, B$ for each of the n residuals $\widehat{\varepsilon}_i$. By using standard bootstrap methods (basic bootstrap) we obtain the n prediction intervals for each of the n pure efficiencies ε_i at the desired level.

We remark that in (B.4) we used the original values of the dependent variable $\widehat{\beta}_\alpha(X_i, Y_i|Z_i)$ to define the original managerial efficiencies. Using the bootstrap values $\beta_\alpha^{*,b}(X_i, Y_i|Z_i)$ obtained in step [2.2] would reproduce the variation of the ε over the n observations, which is not what is needed here. We keep fixed the point of interest $\widehat{\beta}_\alpha(X_i, Y_i|Z_i)$. The bootstrap reproduces the sampling variability due to the estimation of the nonparametric model, the point we evaluate does not move (see Section 5 in Simar and Wilson, 2000 for a detailed discussion in a similar setup).

References

[1] Badin, L., Daraio, C. and L. Simar (2012), How to measure the impact of environmental factors in a nonparametric production model *European Journal of Operational Research*, 223, 818–833.

[2] Bastedo, M. N., Bowman, N. A. (2010), The U.S. News and World Report college rankings: Modeling institutional effects on organizational reputation, *American Journal of Education*, 116, 163–184.

[3] Bonaccorsi A., ed. (2014), *Knowledge, Diversity and Performance in European Higher Education*, Cheltenham, Edward Elgar.

[4] Bonaccorsi A., Daraio C. eds., (2007), *Universities and Strategic Knowledge Creation. Specialization and Performance in Europe*, Edward Elgar Publisher, Cheltenham (UK).

[5] Bonaccorsi A., Daraio C. (2008), The differentiation of the strategic profile of higher education institutions. New positioning indicators based on microdata, *Scientometrics*, 74 (1), 15–37.

[6] Bonaccorsi A., Daraio C., Simar L. (2014), Efficiency and economies of scale and scope in European universities. A directional distance approach, Technical Report n. 8/2014, Sapienza University of Rome.

[7] Bornmann L., Lutz R., Daniel H.D. (2013), Multilevel-Statistical Reformulation of Citation-Based University Rankings: The Leiden Ranking 2011/2012, *Journal of the American Society for Information Science and Technology*, 64(8), 1649–1658.

[8] Bornmann, L. de Moya Anegon, F., Mutz, R. (2013), Do universities or research institutions with a specific subject profile have an advantage or a disadvantage in institutional rankings? A latent class analysis with data from the SCImago ranking, *Journal of the American Society of Information Science and Technology*, 64 (11), 2310-2316.

[9] Bowden, R. (2000) Fantasy Higher Education: University and College League Tables, *Quality in Higher Education*, 6 (1), 41–60.

[10] Buela-Casal, G, Gutierrez-Martinez, O, Bermudez-Sanchez, M Vadillo-Mugnoz, O (2007), Comparative study of international academic rankings of universities, *Scientometrics*, 71 (3), 349–65.

[11] Chambers, R. G., Y. Chung, and R. Färe (1996), Benefit and distance functions, *Journal of Economic Theory*, 70, 407–419.

[12] Cremonini, L., Westerheijden, D.F., Enders, J. (2008), Disseminating the right information to the right audience: Cultural determinants in the use (and misuse) of rankings, *Higher Education*, 55, 373–385.

[13] Daghbashyan Z., Deiaco E., McKelvey M. (2014) How and why does cost efficiency of universities differ across European countries? An explorative attempt using new microdata. In Bonaccorsi A. (ed.), *Knowledge, Diversity and Performance in European Higher Education*, Cheltenham, Edward Elgar.

[14] Daraio C., Bonaccorsi A. et al. (2011), The European University landscape: A micro characterization based on evidence from the Aquameth project, *Research Policy*, 40, 148–164.

[15] Daraio, C. and L. Simar (2007), *Advanced Robust and Nonparametric Methods in Efficiency Analysis. Methodology and applications*, Springer, New York.

[16] Daraio C., Simar L. (2014), Directional distances and their robust versions: Computational and testing issues, *European Journal of Operational Research*, 237, 358–369.

[17] Deprins, D., Simar, L. and H. Tulkens (1984), Measuring labor inefficiency in post offices. In *The Performance of Public Enterprises: Concepts and measurements*. M. Marchand, P. Pestieau and H. Tulkens (eds.), Amsterdam, North-Holland, 243–267.

[18] Docampo D. (2012), Adjusted sum of institutional scores as an indicator of the presence of university systems in the ARWU ranking, *Scientometrics*, 90, 701–713.

[19] Efron, B. and R.J. Tibshirani (1993), *An Introduction to the Bootstrap*. Chapman and Hall, London.

[20] Färe, R., S. Grosskopf and D. Margaritis (2008), Efficiency and Productivity: Malmquist and More, in *The Measurement of Productive Efficiency*, 2nd Edition, Harold Fried, C.A. Knox Lovell and Shelton Schmidt, editors, Oxford University Press.

[21] Farrell, M.J. (1957), The measurement of productive efficiency, *Journal of the Royal Statistical Society*, A(120), 253–281.

[22] Florian, R.V. (2007), Irreproducibility of the results of the Shanghai academic ranking of world universities, *Scientometrics*, 72(1), 25–32.

[23] Goldstein H., Spiegelhalter D.J. (1996), League tables and their limitations. Statistical issues in comparisons of institutional performance, *Journal of the Royal Statistical Society. Series A*, 159(3), 385–443.

[24] Grosskopf, S., Hayes, K., Taylor, L. L. (2014), Applied efficiency analysis in education, *Economics and Business Letters*, 3(1), 19–26.

[25] Hazelkorn E. (2007), The Impact of League Tables and Ranking Systems on Higher Education Decision Making, *Higher Education Management and Policy*, 19 (2), 89–112.

[26] Hazelkorn, E. (2009), Rankings and the battle for world-class excellence: Institutional strategies and policy choices, *Higher Education Management and Policy*, 21(1), 1–22.

[27] Hemlin, S. (1996). Research on research evaluations, *Social Epistemology*, 10(2), 209–250.

[28] Johnes, J. (2006), Measuring teaching efficiency in higher education: an application of data envelopment analysis to economics graduates from UK Universities 1993, *European Journal of Operational Research*, 174(1), 443-56.

[29] Leibenstein H. (1966), Allocative Efficiency vs. "X-Efficiency", *American Economic Review*, 56, 392–415.

[30] Liu, N.C. (2009), The Story of Academic Ranking of World Universities, *International Higher Education*, 54, 2–3.

[31] Liu N.C., Cheng Y., Lin L.(2005), Academic ranking of world universities using scientometrics. A comment to the Fatal Attraction, *Scientometrics*, 64 (1), 101–109.

[32] Liu, N.C., and Y. Cheng (2005), The Academic Ranking of World Universities, *Higher Education in Europe*, 30(2), 127–136.

[33] Lubrano M. (2009), A statistical approach to rankings: Some figures and explanations for European universities. In Dehon C., Jacobs D., Vermandele C. (eds.) *Ranking universities*. Brussels, Editions de l'Universit de Bruxelles.

[34] Marginson, S. (2007), Global position and position-taking: The case of Australia, *Journal of Studies in International Education*, 11(1), 5–32.

[35] Marginson, S., van der Wende M. (2007), To rank or to be ranked: The impact of global rankings in higher education, *Journal of Studies in International Education*, 11(3-4), 306–329.

[36] Moed, H F, W J Burger, J G Frankfort and A F J van Raan (1985), The use of bibliometric data for the measurement of university research performance, *Research Policy*, 14, 131–149.

[37] Porter, S.R., Toutkoushian R.K. (2006), Institutional research productivity and the connection to average student quality and overall reputation. Economics of Education Review 25(6), 605–617.

[38] Provan, D., Abercromby K. (2000), University League Tables and Rankings, *Research in Higher Education*, 45 (5), 443–461.

[39] Saisana, M., D'Hombres B., Saltelli A. (2011), Rickety numbers: Volatility of university rankings and policy implications, *Research Policy*, 40, 165–177.

[40] Safon V. (2013), What do global university rankings really measure? The search for the X factor and the X entity, *Scientometrics*, 97, 223–244.

[41] Salmi J. (2009), *The challenge of establishing world-class universities*, Washington, The World Bank.

[42] Salmi J., Saroyan A. (2007), League Tables as Policy Instruments: Uses and Misuses, *Higher Education Management and Policy*, 19 (2), 33–70.

[43] Sarrico, C. S., Dyson, R. G. (2000), Using DEA for planning in UK universities-an institutional perspective, *Journal of the Operational Research Society*, 51(7), 789–800.

[44] Sarrico, C. S., Teixeira, P. N., Rosa, M. J., Cardoso, M. F. (2009), Subject mix and productivity in Portuguese universities *European Journal of Operational Research*, 197(1), 287–295.

[45] Simar L. and P.Wilson (2000), Statistical Inference in Nonparametric Frontier Models: The State of the Art, *Journal of Productivity Analysis*, 13, 49–78.

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

[47] Simar, L. and P.W. Wilson (2010), Inference From Cross-Sectional Stochastic Frontier Models. *Econometric Review*, 29, 1, 62–98.

[48] Simar, L. and P.W. Wilson (2014), Statistical Approaches for Nonparametric Frontier Models: A Guided Tour, in press, *International Statistical Review*, DP, ISBA, UCL, 2013/47.

[49] Stake, J. E. (2006), The interplay between law school rankings, reputations, and resource allocations: Ways rankings mislead, *Indiana Law Journal*, 82, 229–270.

[50] Toutkoushian, R. K., Webber, K. (2011), Measuring the Research Performance of Post-secondary Institutions. In J. C. Shin, R. K. Toutkoushian, U. Teichler (Eds.), *University rankings. Theoretical basis, methodology and impacts on global higher education*, Dordrecht: Springer Science.

[51] Van Dyke, N. (2005), Twenty years of university report cards, *Higher Education in Europe*, 30(2), 103–125.

[52] Van Leeuwen, T, M Visser, H Moed, T Nederhof and A van Raan (2003), The holy grail of science policy: exploring and combining bibliometric tools in search of scientific excellence, *Scientometrics*, 57, 257–280.

[53] Van Raan A.F.J. (2005), Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods, *Scientometrics*, 62, (1), 133–143.