European Journal of Operational Research 244 (2015) 918–930

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Rankings and university performance: A conditional multidimensional approach ☆

Cinzia Daraio [a,*], Andrea Bonaccorsi [b], Léopold Simar [c]

[a] Department of Computer, Control and Management Engineering Antonio Ruberti (DIAG), University of Rome "La Sapienza", Via Ariosto 25, Rome 00185, Italy
[b] Dipartimento di Ingegneria dell'Energia, dei Sistemi, del Territorio e delle costruzioni (DESTEC), University of Pisa, Italy and National Agency for the Evaluation of Universities and Research Institutes (ANVUR)
[c] ISBA, Université Catholique de Louvain, Louvain-la-Neuve, Belgium and DIAG University of Rome "La Sapienza", Italy

ARTICLE INFO

Article history:
Received 18 June 2014
Accepted 2 February 2015
Available online 9 February 2015

Keywords:
Rankings
European universities
Conditional directional distances
Robust frontiers
Bootstrap

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.

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 (European Universities Micro Data) census of Higher

☆ Research supported by the University of Rome "La Sapienza", "Progetto di Ateneo 2013" and IAP Research Network P7/06 of the Belgian State (Belgian Science Policy). The opinions expressed in this paper are personal and do not involve the institution of the second author (ANVUR).

* Corresponding author. Tel.: +39 06 77274 068, +393488709425; fax: +39 06 77274 074.

E-mail addresses: daraio@dis.uniroma1.it (C. Daraio), a.bonaccorsi@gmail.com (A. Bonaccorsi), leopold.simar@uclouvain.be (L. Simar).

http://dx.doi.org/10.1016/j.ejor.2015.02.005
0377-2217/© 2015 Elsevier B.V. All rights reserved.

Education Institutions (HEIs) 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 B), 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 Social Sciences and Humanities (SSH), 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 non-academic 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 & Dyson, 2000; Sarrico, Teixeira, Rosa, & Cardoso, 2009 and Grosskopf, Hayes, & Taylor, 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, Daraio, Bonaccorsi, and Simar (2015) propose a robust directional distance approach to analyze economies of scale and specialization 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 taking into account also the volume of scientific production and the teaching realized. Research quality is hence the main output of interest. It is measured by a factor built taking into account international collaborations, normalized impact of research, high quality and excellence rate of publications. By applying a robust directional distance technique, we consider as non-discretionary outputs the volume of teaching and research carried out as well. 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 analyzed. Section 4 illustrates the methodology and is complemented by Appendix B. Finally, Section 5 presents the main results, while Section 6 concludes the paper.

2. University rankings: a guided tour of the critical literature

In this section we present the main lines of criticism to university rankings in the four chapters anticipated in Section 1. Other classifications are certainly possible. For the sake of clarity, criticisms classified in categories (a, monodimensionality) and (d, lack of input–output structure) deal with the substantive content of rankings, i.e., the data included (or missing), while studies under (b, statistical robustness) and (c, dependence on size and subject mix) 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 Social Sciences and Humanities 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 focus only on research 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. Therefore, there is a demand for including information on teaching as well as research outputs of universities. Existing rankings include only a small set of indicators, whose meaning in terms of overall education activity of universities is questionable: the Alumni Nobel and Field prizes (10 percent) in ARWU (Academic Ranking of World Universities), student/staff ratios (20 percent weight), international students (5 percent) and international staff (5 percent) in QS (Quacquarelli Symonds) World University Rankings, and income per academic (2.25 percent), undergraduates admitted per academic (4.5 percent), ratio of international to domestic students (2.5 percent), ratio of international to domestic staff (2.5 percent) in THE (Times Higher Education Rankings). These proxies are considered unreliable and highly volatile by most analysts, as it is witnessed by the lack of consistency across various rankings, with the exception of the few top positions (Saisana, D'Hombres, & Saltelli, 2011; Salmi & Saroyan, 2007).

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, Westerheijden, & Enders, 2008). The over reliance on research indicators may induce biased decisions (Bastedo & 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 toward 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, Burger, Frankfort, & van Raan, 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, a university's completion and degree of achievement are strongly correlated with the quality of students it takes in.

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 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; Florian, 2007; Van Dyke, 2005).

One line of reasoning has stressed the importance of not using just one ranking but multiple ones. More generally, Van Leeuwen, Visser, Moed, Nederhof, and van Raan (2003) have underlined the importance of 'using multiple indicators instead of only one' (Van Leeuwen 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, & 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 & 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 & Spiegelhalter, 1996, p. 405).

A third direction 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 & 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 Supplement rankings, and that for a certain number 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 disciplines (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 & 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 wider variety 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, Gutierrez-Martinez, Bermudez-Sanchez, & Vadillo-Mugnoz, 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 & 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 (Organisation for Economic Cooperation and Development) 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 & 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 percent 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 ask 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 et al. (2008), who argue

[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, Deiaco, and McKelvey (2014).

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 toward research, and reputation" (p. 238). As it is clear from this list, only a few of these factors, such as orientation toward 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 at the international level is the one of efficiency, or the relation between input and output.[2] The joint consideration of outputs and resources employed is the starting point for university strategy (Bonaccorsi & Daraio, 2007) and for the positioning of universities with respect to their peers (Bonaccorsi & Daraio, 2008).

Table 1
Definition of inputs, outputs and conditioning factors.

Input/output/conditioning factor	Definition
Input	
NACSTA (x_1)	Number of non-academic staff
ACSTAF (x_2)	Number of academic staff
PEREXP (x_3)	Personnel expenditures (PPS)
NOPEXP (x_4)	Non-personnel expenditures (PPS)
FINP	Input factor including: NACSTA, ACSTAF, PEREXP, NOPEXP
Output	
TODEG5 (y_1)	Total degrees ISCED 5
TODEG6 (y_2)	Total degrees ISCED 6 (Doctorate)
PUB (y_3)	Number of published papers (Scimago)
IC (y_4)	International collaboration (Scimago)
NI (y_5)	Normalized impact (Scimago)
Q1 (y_6)	High quality publications (Scimago)
EXC (y_7)	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.

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 (Directorate General for Education and Culture), DG RTD (Directorate General for Research and Innovation), 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 & 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 postgraduate 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 2005–2009) 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.[3] 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 and 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), percent 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), percent of publications that a university publishes in the first quartile (25 percent) in their categories as ordered by Scimago Journal Rank indicator.
- Normalized Impact (NI), in percent shows the relationship between a university's average scientific impact and the world average set to a score of 1.
- Excellence Rate (EXC), percent of university output that is included in the 10 percent 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. The choice of these variables has been carried out by making a compromise between relevance of the factors and availability of data. For instance, capital expenditures would be an interesting input to include in the analysis but unfortunately there were not available data.[4]

[2] This topic has been addressed in national contexts by various contributions (see e.g. Abbott & Doucouliagos, 2003; Flegg, Allen, Field, & Thurlow, 2004; Johnes, 2008, 2013; Worthington & Lee, 2008).

[3] The integration has been carried out within the Smart.CI.EU (Sapienza microdata architecture for education, research and technology studies. A Competence-based data Infrastructure on European Universities), an experimental data infrastructure created within a research project funded by Sapienza University of Rome and owned at the Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome.

[4] As a consequence, the omission of capital expenditure might cause possible distortion in the comparison of university performance. A factorial analysis has been done on the inputs listed in Table 1 (NACSTA, ACSTAF, PEREXP, NOPEX) and on the base of its results (see Appendix A) an input factor was calculated for the empirical investigation.

Table 2
Descriptive statistics.

Variable	25th perc.	Median	Average	75th perc.	Std.
NACSTA	562	1040	1497	1807	1408
ACSTAF	687	1164	1470	1970	1058
PEREXP	54 714 812	103 370 360	142 577 883	187 468 894	121662 902
NOPEXP	27 258 575	58 097 154	87 111 330	100 277 918	94 924 980
TODEG5	1750	3205	3882	4985	3146
TODEG6	54	121	201	275	214
PUB	1515	3609	5571	7530	5626
IC	33	38	39	44	9
NI	1.10	1.30	1.30	1.50	0.31
Q1	44	53	51	60	13
EXC	11	15	15	20	6
SIZE	10 056	16 755	20 258	24 550	17 486
SPEC	0.60	0.70	0.69	0.80	0.13

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.

Table 2 reports some descriptive statistics (25th percentile, median, average, 75th percentile and standard deviation) on the sample that will be analyzed in the paper. It would have been interesting also to consider in the analysis other relevant variables, such as external research funding of universities. Unfortunately, this information was not available and hence has not been included in the analyses.[5]

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 & 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. $H_{XY}(x, y) = \text{Prob}(X \le x, Y \ge y)$.

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, Chung, and Färe (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 & Simar, 2014[6]):

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

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. As we shall also discuss later, 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.).

Consistent nonparametric estimators of Eq. (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 (see Appendix A for further details) and given the high correlations observed among variables 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, Grosskopf, and Margaritis (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 analyzed 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

[5] A potentially interesting line of future research could consist in formulating a network problem, either in terms of reallocation within university systems or thinking about actual success of graduates as ultimate outcome of interest (see e.g. Grosskopf, Hayes, Taylor, & Weber, 2012).

[6] See the references cited there.

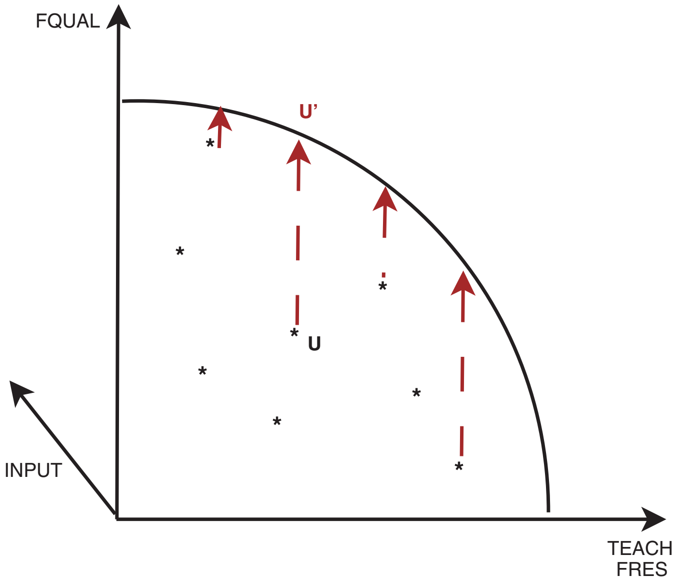

Fig. 1. The production model of excellent science by European universities: a simplified illustration.

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 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.

It is worth noting that in our case, by using only one discretionary output (all others are non-discretionary and their directions are set to zeros) the resulting ranks are independent of the direction value. That is, since only one direction is active, it does not matter (at a scaling factor) which direction we choose, e.g. the mean, the median and so on, this has no influence on the resulting rankings that will keep unchanged. Nevertheless, if we would add any other "active" direction (asking one of our non-discretionary output to become active), then the direction will play a role and the obtained ranking would probably be different. The flexibility of directional distances is that they allow us to model the phenomenon as we want. In our case, it is an output direction that focuses only on quality of research keeping the other outputs as fixed. Other modeling strategies could be chosen and this is thanks to directional distances flexibility. Moreover, a nice property of a constant (e.g. unitary, mean, median) direction vector is that it greatly facilitates aggregation.[7]

See Fig. 1 for an illustration. In Fig. 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 toward u'.

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 us 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:

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

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. These latter are two of the most important sources of heterogeneity of university performance. Our goal is 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.[8]

5. Accounting for statistical robustness

   One of the main methodological issues 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 B).

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 & 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 at the support of the distribution H_{XY} we benchmark the unit against a point which leaves on average $\alpha \times 100$ percent 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.

As shown in Daouia, Simar, and Wilson (2014), the directional distance and its estimate, by contrast with the radial measure, have the desired property of always being monotonically increasing with the input variables (the inefficiency score increases when the inputs increase, all other variables being fixed), see details in Daouia et al. (2014).

[7] We did not exploit this property in this paper, but consider it as an interesting topic for further research, along the lines of Färe, Grosskopf, and Primont (2007).

[8] However, also other variables could be considered, such as the localization of universities. The investigation of this variable effect is left for future research.

A nonparametric estimator[9] can be found in Daraio and Simar (2014) which detail the case with some non-discretionary inputs and/or outputs that we apply in this paper.

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

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

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 values: 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 & 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 Appendix A.

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.[10]

The idea is very simple, we only have to replace $H_{XY}(x, y)$ in the above unconditional model by $H_{XY|Z}(x, y|Z = z) = \text{Prob}(X \le x, Y \ge y|Z = z)$ 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 nonparametric estimator of $H_{XY|Z}(x, y|Z = z)$ when some directions are set to zero as well as its robust version[11] that we apply in this paper.

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 & 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 require 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 & 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

$\mu(z) = \mathbb{E}(\beta(X, Y; d_x, d_y|Z = z))$ and

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

These two functions can be estimated nonparametrically from a sample of observations $\{Z_i, \widehat{\beta}(X_i, Y_i; d_x, d_y|Z_i)\}$, $i = 1, \ldots, n$ by using, e.g., Nadaraya–Watson or local linear estimates (see Daraio & 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 (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 us 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.

As explained above, if we want to make an analysis which is robust to extreme and outlying data points, it is preferable to use the robust version of the efficiency scores $\beta_{\alpha}(x, y; d_x, d_y)$ selecting a value of α near 1 to provide a robust version of the full frontier. In addition, by using the nonparametric estimator of these α-efficiency scores, we can build bootstrap confidence bounds for the resulting managerial efficiency scores, as described in details in Appendix B. The idea is to adapt the bootstrap algorithm provided in Daraio and Simar (2014) to our setup here.

[9] Denoted in Section 5 $\beta_{\alpha,\text{FDH}}$ because it is the robust version of a directional distance based on a nonconvex FDH (Free Disposal Hull, Deprins, Simar, & Tulkens, 1984) estimator.

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

[11] Denoted in Section 5 $\beta_{\alpha,\text{FDH}|Z}$.

Table 3
Efficiency results: averages by country.

Country	Country code	Coverage (in percent)	#obs	#dom	$\widehat{H}_{XY}$	$\beta_{\alpha,\mathrm{FDH}}$	$\beta_{\alpha,\mathrm{FDH}\vert Z}$
Austria	AT	88.13	13	4.46	0.0111	0.049757	0.080548
Belgium	BE	56.32	3	3.33	0.0083	0.028793	0.083364
Switzerland	CH	99.25	10	1.20	0.0030	−0.105144	0.009617
Czech Republic	CZ	81.71	11	3.55	0.0088	0.201562	0.196404
Germany	DE	95.01	62	11.94	0.0298	0.225836	0.239438
Spain	ES	100	42	6.76	0.0169	0.181811	0.172416
Finland	FI	91.68	5	2.80	0.0070	0.115628	0.147933
Hungary	HU	63.11	6	27.50	0.0686	0.354777	0.371831
Ireland	IE	72.60	6	3.33	0.0083	0.023005	0.104317
Italy	IT	93.26	45	5.31	0.0132	0.096320	0.142867
Netherlands	NL	98.25	7	5.57	0.0139	0.085189	0.130575
Norway	NO	85.91	8	6.25	0.0156	0.140668	0.145187
Romania	RO	67.75	7	2.71	0.0068	0.171315	0.255163
Sweden	SE	93.43	9	2.33	0.0058	−0.077888	0.055961
United Kingdom	UK	84.52	73	1.97	0.0049	−0.067431	0.063181
All sample	EU	87.98	313	6.07	0.0151	0.092477	0.145114

Notes: Only countries with at least 3 observations are reported in the table. The coverage is calculated with respect to the total number of academic staff. The last line reports the average over the whole analyzed sample.

5. Results

In this section we report the main results obtained from our analysis on the European university dataset described in Section 3. We estimated individual efficiency scores for each university in our sample but present the results by averaging these individual scores at the country level to facilitate the interpretation.

Table 3 reports in the columns: Country, country code, coverage in percentage, 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 ($\beta_{\alpha,\mathrm{FDH}}$) and robust directional measure of efficiency conditioned to SIZE and SPEC, our Z variables ($\beta_{\alpha,\mathrm{FDH}|Z}$). The coverage, expressed in percentage, has been calculated with respect to the total academic staff. We did the sum of the total academic staff of all the universities in our sample by country, and made the proportion of this value with the sum of the total academic staff of all the universities doctorate awarding available at country level in the EUMIDA DC2 database. We observe that the coverage of our sample with respect to all the universities doctorate awarding in the respective European countries is very good as it goes from 56.32 percent for Belgium, to 100 percent for Spain. Overall, the sample coverage is of 87.98 percent. Nevertheless, we suggest to take into account the existing differences in country coverage for the following reading and interpretation of the results.

The last line of the table shows the average on the overall sample. 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, Sweden and Switzerland, followed by Belgium, Austria, Ireland and Netherlands, this appears if we consider both the unconditional efficiency scores ($\beta_{\alpha,\mathrm{FDH}}$) and the conditional efficiency scores ($\beta_{\alpha,\mathrm{FDH}|Z}$).

Table 4 reports the estimated gaps in percentage of the outputs produced by the units. As expected, countries that perform better on average are again: UK, Sweden and Switzerland, followed by Belgium, Austria, Ireland and Netherlands. Now also Norway performs slightly higher than the European average.

We provide an interpretation of these results in the concluding section (Section 6), advancing a conjecture that of course should be further investigated.

Fig. 2 illustrates the distribution of the Managerial efficiency scores estimated over the whole European sample. We remind that the value

Table 4
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

Notes: only countries with at least 3 observations are reported in the table. For #DEG5, #DEG6 and #PUB we have all zero gaps because these variables are considered non-discretionary and their directions are set to zero. The last line reports the average over the whole analyzed sample.

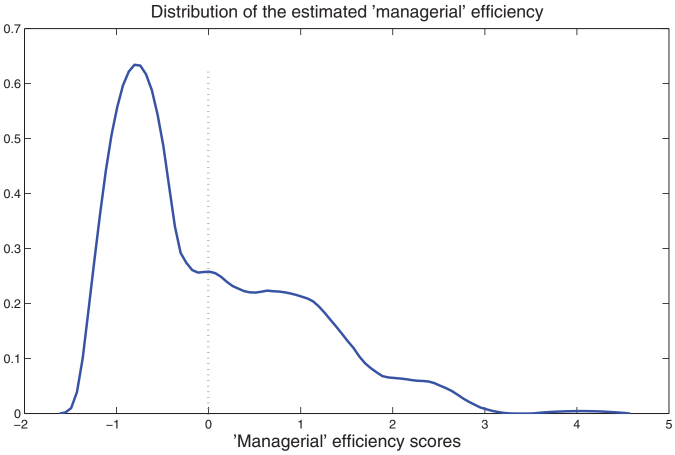

Fig. 2. Nonparametric kernel distribution of the estimated managerial efficiency scores.

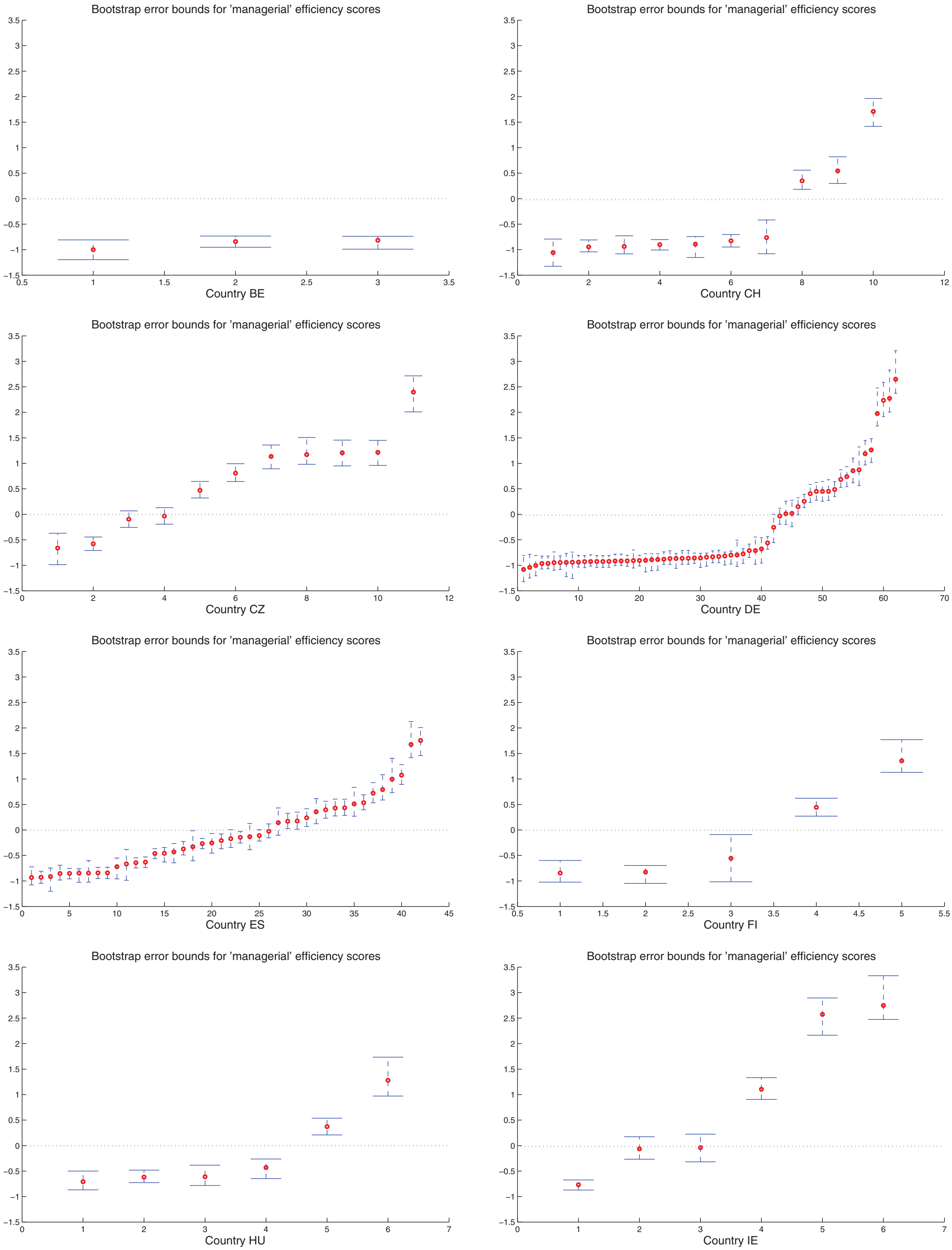

Fig. 3. Managerial efficiency scores by country with bootstrap error bounds.

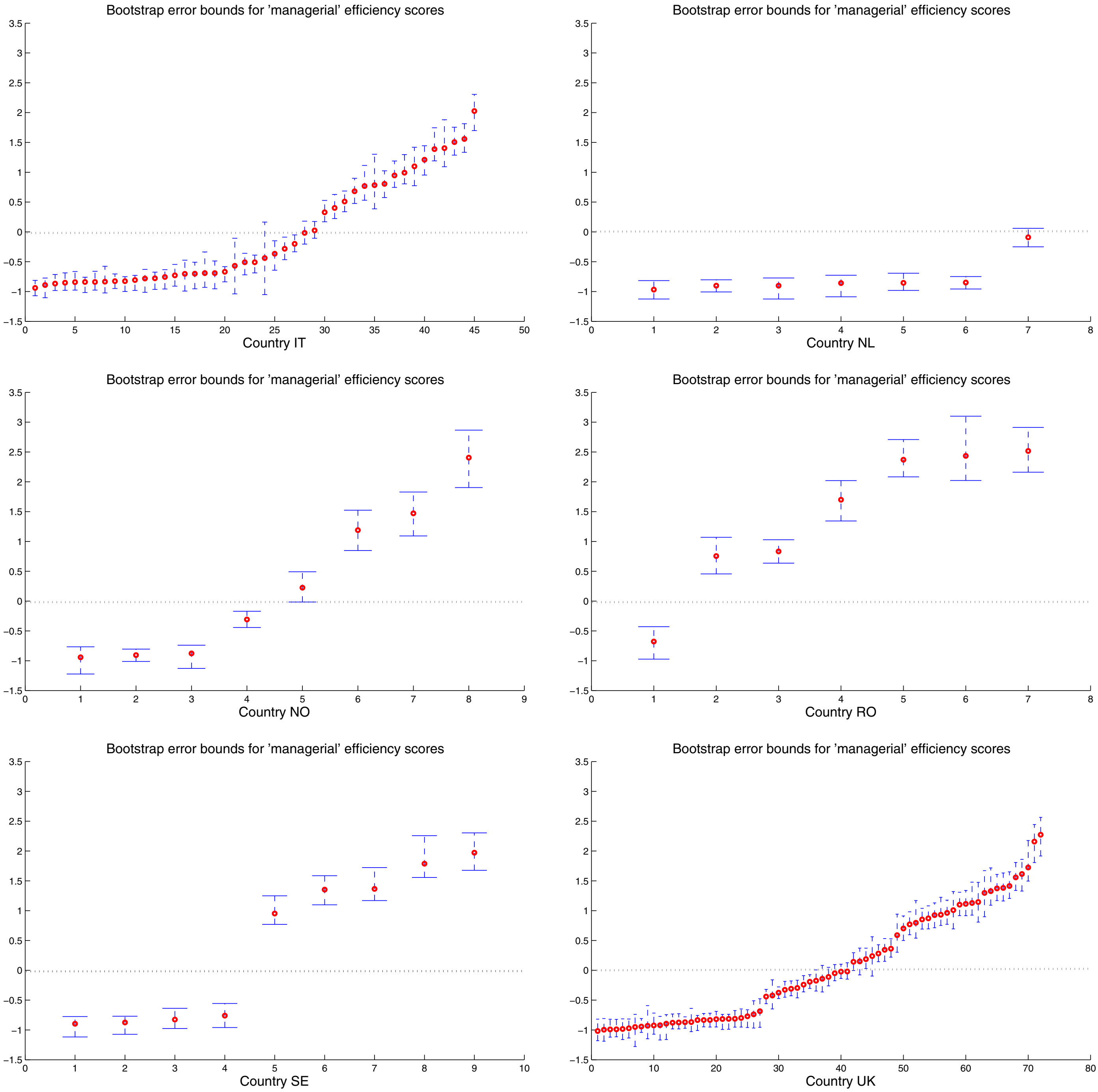

Fig. 4. Managerial efficiency scores by country with bootstrap error bounds. Cont.

of managerial efficiency equal to zero corresponds to the European average. 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 is a non-standard distribution, different from the commonly assumed distributions (such as exponential or half normal) in parametric frontier approaches. It shows a peak around −0.9, a value that indicates very good managerial performance. Besides there is a long queue of universities with worst performance. A detailed view on the managerial efficiency scores of individual universities by country is offered in Figs. 3 and 4.

The graphics reported in Figs. 3 and 4 illustrate the estimated managerial efficiency scores for each university by country, providing their relative bootstrap error bounds.

These figures clearly 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 Figs. 3 and 4 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.

6. Discussion of results and 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 result 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 results provide preliminary evidence on how to disentangle different dimensions of performance (education, volume of research, quality of research) in the attempt of identifying directions 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 recognized. 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 with 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 might follow 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 toward 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 (Bonaccorsi & Daraio, 2007; Daraio et al., 2011).

Appendix A. Factorial analysis and gaps calculation

As described in Section 4, in the empirical analysis for the inputs, we replace the 4 scaled inputs by their best (non-centered) linear combination, defined as FINP, as described in Table 1. In so doing, we control the information that we lose in aggregating the variables, that should not be too high, say less than 15 percent. We also control the correlation of the resulting univariate input factor with the 4 original inputs, that should be high.

The obtained results are the following: $\text{FINP} = 0.48x_1 + 0.56x_2 + 0.52x_3 + 0.44x_4$, where we see that the factor is a weighted average of the 4 inputs. FINP explains 94 percent of total inertia of original data (correlations of the FINP with the original inputs are 0.93, 0.91, 0.98, 0.92). We follow the same procedure with the outputs. The results for the two factors are: $\text{FRES} = 0.70y_2 + 0.71y_3$ and $\text{FQUAL} = 0.56y_4 + 0.51_y5 + 0.56_y6 + 0.33_y7$, where FRES and FQUAL are defined in Table 1. FRES explains 96 percent of total inertia of original data (correlations of the FRES with the original data are 0.96 and 0.96), while FQUAL explains 98 percent of total inertia of original data (correlations of FQUAL with original values are 0.7, 0.9, 0.9, 0.9).

Therefore, in the analysis, the factor $F_{\widetilde{X}}$, where $\widetilde{X}$ denotes the matrix of the selected inputs to be aggregated, will act as a single observed input and will be combined with the outputs (or other output factors) along the lines of the techniques developed above. The gaps obtained from this analysis, denoted by G_F, are thus expressed in the units of the factors $F_{\widetilde{X}}$ used and not in the units of the original variables included in $\widetilde{X}$. We know that the value of the input factor variable on the efficient frontier is given by $\widehat{F}^{\partial}_{\widetilde{X}} = F_{\widetilde{X}} + G_F$. It is easy to check that the coordinates of $F_{\widetilde{X}}$ in the original units of $\widetilde{X}$ are given by $F_{\widetilde{X}}a'_1$, where a_1 is the eigenvector of $\widetilde{X}'\widetilde{X}$ corresponding to its largest eigenvalue. For the same reason, the coordinates of the frontier points are $\widehat{F}^{\partial}_{\widetilde{X}}a'_1$, so the measure of the gaps in the units of $\widetilde{X}$ are given by $G_{\widetilde{X}} = G_F a'_1$.

The same procedure applies to the other output factors calculated and used in the analysis.

Appendix B. Bootstrap error bounds for managerial efficiency scores

We use standard bootstrap methods (for an introduction, see Efron & 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 percent) 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:

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

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 & Wilson, 2000 for a detailed discussion in a similar setup).

References

Abbott, M., & Doucouliagos, H. (2003). The efficiency of Australian universities: A data envelopment analysis. *Economics of Education Review*, *22*, 89–97.

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

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.

Bonaccorsi, A. (Ed.) (2014). *Knowledge, diversity and performance in European higher education*. Cheltenham: Edward Elgar.

Bonaccorsi, A., & Daraio, C. (Eds.) (2007). *Universities and strategic knowledge creation. Specialization and performance in Europe*. Cheltenham, UK: Edward Elgar Publisher.

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.

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.

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.

Bowden, R. (2000). Fantasy higher education: University and college league tables. *Quality in Higher Education*, *6*(1), 41–60.

Buela-Casal, G., Gutierrez-Martinez, O., Bermudez-Sanchez, M., & Vadillo-Mugnoz, O. (2007). Comparative study of international academic rankings of universities. *Scientometrics*, *71*(3), 349–365.

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

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.

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.

Daouia, A., Simar, L., & Wilson, P. W. (2014). Measuring firm performance using nonparametric quantile-type distances. *Econometric Review* (forthcoming).

Daraio, C., & Bonaccorsi, A. (2011). The European university landscape: A micro characterization based on evidence from the Aquameth project. *Research Policy*, *40*, 148–164.

Daraio, C., Bonaccorsi, A., & Simar, L. (2015). Efficiency and economies of scale and specialization in European universities: A directional distance approach (under review for the Journal of Informetrics).

Daraio, C., & Simar, L. (2007). *Advanced robust and nonparametric methods in efficiency analysis. Methodology and applications*, Springer, New York.

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

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

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.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*, London: Chapman and Hall.

Färe, R., Grosskopf, S., & Margaritis, D. (2008). Efficiency and productivity: Malmquist and More. In H. Fried, C. A. Knox Lovell, & S. Schmidt (Eds.), *The measurement of productive efficiency* (2nd ed.). Oxford University Press.

Färe, R., Grosskopf, S., & Primont, D. (Eds.) (2007). *Aggregation, Efficiency and Measurement*. LCC, New York: Springer Science and Business Media.

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

Flegg, T., Allen, D., Field, K., & Thurlow, T. W. (2004). Measuring the efficiency of British universities: A multi-period data envelopment analysis. *Education Economics*, *12*(3), 231–249.

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

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.

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

Grosskopf, S., Hayes, K., Taylor, L. L., & Weber, W. (2012). Centralized or decentralized control of school resources? A network model. *Journal of Productivity Analysis*, 1–12 10.1007/s11123-013-0379-2.

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.

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.

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

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), 443–456.

Johnes, J. (2008). Efficiency and productivity change in the English higher education sector from 1996/97 to 2004/05. *The Manchester School*, *76*(6), 653–674.

Johnes, J. (2013). Efficiency and mergers in english higher education 1996/97 to 2008/9: Parametric and non-parametric estimation of the multi-input multi-output distance function. *The Manchester School*, *82*(4), 465–487.

Leibenstein, H. (1966). Allocative efficiency vs. "x-efficiency". *American Economic Review*, *56*, 392–415.

Liu, N. C. (2009). The story of academic ranking of world universities. *International Higher Education*, *54*, 2–3.

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.

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'Université de Bruxelles.

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

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.

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

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.

Provan, D., & Abercromby, K. (2000). University league tables and rankings. *Research in Higher Education*, *45*(5), 443–461.

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

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

Salmi, J., & Saroyan, A. (2007). League tables as policy instruments: Uses and misuses. *Higher Education Management and Policy*, *19*(2), 33–70.

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.

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.

Simar, L., & Wilson, P. (2000). Statistical inference in nonparametric frontier models: The state of the art. *Journal of Productivity Analysis*, *13*, 49–78.

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

Simar, L., & Wilson, P. W. (2010). Inference from cross-sectional stochastic frontier models. *Econometric Review*, *29*(1), 62–98.

Simar, L., & Wilson, P. W. (2014). Statistical approaches for non-parametric frontier models: A guided tour. *International Statistical Review*, 1–34 doi:10.1111/insr.12056.

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

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.

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

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

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.

Worthington, A. C., & Lee, B. L. (2008). Efficiency, technology and productivity change in Australian universities 1998–2003. *Economics of Education Review*, *27*(3), 285–298.