CESifo
WORKING
PAPERS

8044
2020

January 2020

Efficiency of European Universities: A Comparison of Peers

Lars Herberholz, Berthold U. Wigger

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 8044*

Efficiency of European Universities: A Comparison of Peers

Abstract

The European higher education landscape has become increasingly integrated causing competition among universities that is no longer bound to national borders. In view of this development, the present paper investigates the relative efficiency of 450 European universities between 2011 and 2014. The novelty of our approach lies in its extended coverage of university outputs and in the thorough peer-group selection process that accounts for high diversity in subject profiles. More specifically, assignment to peer-groups builds on proximity in subject space to ensure valid comparisons between universities. Exploring potential efficiency drivers, we uncover considerable effect heterogeneity between subject clusters, which is indicative of distinct technologies and calls for carefully designed policy measures. Yet institutional size and the ability to seek external funding are largely identified to be primary efficiency drivers.

JEL-Codes: H520, I230, I280.

Keywords: university efficiency, peer selection, data envelopment analysis, clustering.

Lars Herberholz
Karlsruhe Institute of Technology
Karlsruhe / Germany
lars.herberholz@kit.edu

Berthold U. Wigger
Karlsruhe Institute of Technology
Karlsruhe / Germany
berthold.wigger@kit.edu

January 13, 2020
This research greatly benefitted from funding by the Federal Ministry of Education and Research (BMBF) under grant number 01PB14008. Furthermore, we are thankful for access to high-performance computing resources provided by the state of Baden-Württemberg through bwHPC. Of course, responsibility for the content lies with the authors alone.

1 Introduction

As a consequence of European integration, universities in Europe are more and more competing for students, research funding, and scientific personnel across borders. It is therefore all the more important not only to assess these universities relative to their national peers, but to consider comparisons on a broader geographical scale. Yet most studies on higher education efficiency have so far confined attention to single countries. Against this background, the present paper adds to the scarce literature on cross-country studies by investigating how efficiently universities from 16 European countries use the resources at their disposal. Having estimated relative efficiency scores, we further aim to identify relevant efficiency drivers, such as funding or personnel structure, by means of regression techniques.

Adopting a European perspective enables us to make a methodological contribution to the literature on university efficiency. Given that input-output patterns notably depend on subject composition, universities are far from being considered homogeneous. We propose to address this issue with clustering methods and individual peer-group selection that both build on distance in subject space. We hereby avoid any kind of unreasonable comparison of, for instance, technical universities and business schools. Such an approach naturally comes with the requirement of a sufficiently large number of institutions, which we meet by using the European Tertiary Education Register (ETER) along with Elsevier's database Scopus. In total, these two data sources allow us to compile detailed information on a comprehensive set of 450 universities.

Overall, our main results can be summarised as follows. First, it becomes evident that efficiency comparisons should only be made for universities with similar subject focus. Otherwise, efficiency scores would be more indicative of subject differences, e.g. higher costs in medical or technical studies relative to social studies, than of more or less efficient resource use. Second, efficiency drivers show substantial effect heterogeneity between subject clusters, which suggests that universities are shaped by different technologies. However, we third provide evidence that third-party funding shares and institutional size are to a large extent efficiency-enhancing.

The remainder of the paper is organised as follows. Section 2 discusses related literature and further elaborates on our key ideas. Section 3 establishes the classification scheme. Section 4 introduces the statistical Data Envelopment Analysis (DEA) approach. Section 5 explores estimated efficiency scores. Section 6 identifies efficiency drivers. Section 7 provides robustness checks. Section 8 briefly concludes.

2 Related Literature

Detecting inefficiencies within educational institutions has attracted much scholarly attention, most notably leading to empirical studies that employ various frontier efficiency techniques. Starting in the 1980s, numerous studies have focused on different types of institutions including primary and secondary schools, universities, and university departments, or countries as a whole. For comprehensive reviews, see e.g. Worthington (2002) or De Witte and López-Torres (2017).

Focussing on higher education, the foundations have been laid by studies that were conducted on a single-country basis. Historically, Anglo-Saxon countries were at the centre of most early frontier analyses. For instance, within the United States, Ahn, Charnes, and Cooper (1988) and Ahn and Seiford (1993) were both concerned with comparing public and private doctoral-granting institutions whereas Breu and Raab (1994) confined attention to the nation's top ranked institutions. Australian universities have also been subject to frequent assessment by Coelli (1996), Avkiran (2001), and Abbott and Doucouliagos (2003), among others. The same applies to institutions in the United Kingdom that have been analysed in depth. While British studies on academic efficiency were first conducted at the department level (e.g. Tomkins and Green, 1988, Beasley, 1990, and Johnes and Johnes, 1995, on accounting, chemistry and physics, and economics departments, respectively), several contributions at the university level were soon to follow, for instance by Athanassopoulos and Shale (1997), Sarrico et al. (1997), and Johnes (2006). Furthermore, McMillan and Datta (1998) provided insights into the relative performance of Canadian universities while Taylor and Harris (2004) addressed this topic with regard to South Africa. Apart from the origins in the Anglo-Saxon area, higher education efficiency has emerged as a (research) topic of global interest as can be seen by studies covering institutions in Austria (Leitner, Prikoszovits, Schaffhauser-Linzatti, Stowasser, & Wagner, 2007), Germany (Kempkes & Pohl, 2010), Italy (Agasisti & Salerno, 2007), Greece (Katharaki & Katharakis, 2010), Brazil (Zoghbi, Rocha & Mattos, 2013), Mexico (Sagarra, Mar-Molinero, & Agasisti, 2017), or China (Johnes & Yu, 2008).

Frontier techniques are essentially driven by the number of (decision-making) units under assessment since efficiency is defined endogenously by the subset of units with the highest productivity. The aforementioned studies are therefore bound to national efficiency frontiers, which is apparently at odds with the widespread view of universities competing on a global scale. The limitations of country-specific studies have indeed motivated a promising stream of literature, i.e. cross-country studies. Among them, Joumady and Ris (2005) were arguably the first to make a contribution by exploiting a postal survey sent to young professionals three years after graduation. In total, 209 institutions from eight European countries were assessed regarding their capacity to prepare students for labour market transition. Due to the unique survey setting, this work constitutes a rather special case. In comparison, most subsequent studies have pursued an alternative path by using administrative data from national agencies, which usually calls for manual adjustment to improve comparability. This might be a reason why several studies started to adopt a two-country perspective. For instance, Agasisti and Johnes (2009) compared universities from England and Italy noting that the latter ones were largely outperformed in the academic year 2003/04. Following a similar methodology, Agasisti and Pérez-Esparrells (2010) conducted an analysis of Italian and Spanish universities. Referring to the academic year 2004/05, Italian universities were this time found to operate at higher efficiency levels. In fact, further two-country studies have been centred around Italy based on data from 2000 onwards. Agasisti and Pohl (2012) observed lower efficiency levels of Italian universities relative to their German counterparts while comparisons to Dutch (Agasisti & Haelermans, 2016) and Polish (Agasisti & Wolszczak-Derlacz, 2016) institutions revealed that efficiency differentials are mostly model-dependant.

Overall, two-country studies can be regarded as a first step to account for increasing internationalization in higher education. However, comparisons on a broader geographical scale are still required to obtain a more complete picture of cross-border competition and production possibilities. Apart from Joumady and Ris (2005), only a handful of studies have addressed this need to date, which, for the most part, can be explained by the lack of comparable micro-data at the institutional level (Wolszczak-Derlacz, 2017). Wolszczak-Derlacz and Parteka (2011) approached this issue by means of a multitude of sources that led to a dataset on 259 universities from seven European countries. Extending the scope of analysis, Wolszczak-Derlacz (2017) compared 152 U.S. to 348 European universities from ten countries, again based on datasets collected manually. Both studies consistently show that efficiency scores tend to vary clearly both within and between countries. Other studies were built on the projects Aquameth and Eumida, which were initial attempts by the European Commission to construct an integrated database on higher education institutions. Exploiting these databases, Daraio, Bonaccorsi, and Simar (2015a) investigated economies of scale and specialization, Bolli et al. (2016) emphasised the role of competitive funding while Daraio, Bonaccorsi, and Simar (2015b) proposed an advanced approach to university rankings by means of frontier techniques. Albeit these recent contributions, cross-country studies on higher education efficiency are evidently still scarce. We therefore aim to extend this strand of literature with the help of a novel dataset. To the best of our knowledge, we are the first to utilise ETER for efficiency purposes, which enables us to expand our scope beyond past research and present a strong case for the validity of our findings.

Apart from providing an extended cross-country perspective on university efficiency, we provide a methodological contribution to a second stream of literature that has rather been neglected in recent work. Specifically, we take the view that subject mix differentials have to be addressed comprehensively to avoid a well-known pitfall within DEA applications, i.e. comparing non-homogeneous units (Dyson et al., 2001). In fact, several studies have highlighted various systematic differences between academic fields. For instance, Tierny (1980) provides early evidence on costs per student at liberal arts colleges and shows that chemistry departments exceed political science departments by up to 100%. The general notion that social sciences incur lower cost levels than physical sciences is also confirmed by Dundar and Lewis (1995), who, additionally, discover the highest costs in the field of engineering sciences. Further cost studies have come to similar conclusions. Zimmerman and Altonji (2018) examine instructional spending in the Florida State University System and discover substantial heterogeneity. According to their results, engineering graduates entail costs that are almost double the amount found in low-cost majors, such as business. Consistently, Filipini and Lepori (2007) explore expenditure levels by Swiss universities and report the highest values for technical sciences along with medicine. In view of the sharp differences between disciplines, they emphasise that cost comparisons of universities are at risk of being distorted if subject composition is left unconsidered. Johnes (1990) adds to this line of reasoning by stating that over two thirds of the variation in unit costs among UK universities is attributable to subject mix alone. There appear to be two main reasons for these patterns. On the one hand, STEM-related fields but also medicine generally require physical resources to a different extend and magnitude, e.g. basic materials, clinical and mechanical equipment, laboratories, and other costly facilities. On the other hand, some of these subjects are considered to be more labour-intensive with higher levels of interaction between students and faculty members, which is reflected by a different personnel structure (Kempkes & Pohl, 2010).[1]

Closely related to the cost dimension is the relevance of external research funding, which is also well documented to be subject-dependant. The findings present a clear picture in the sense that STEM-related fields along with medicine are most engaged in third-party collaborations. Social sciences and humanities, on the contrary, are clearly underrepresented, likely due to less commercial potential in these fields (Bonaccorsi, Secondi, Setteducati, & Ancaiani, 2014; Gulbrandsen & Smeby, 2005; Hornbostel, 2001). Relying on third-party funding as a proxy for research performance, as often favoured in the absence of bibliometric data, therefore becomes a two-fold problem. Not only does this practice raise economic concerns about confounding inputs with outputs (Johnes & Johnes, 1995), but it also introduces unfair judgement. Publication and citation counts provide preferable output measures are, however, prone to bias, too. Shin and

[1] The latter argument is presumably not restricted to STEM-subjects and medicine. For instance, music and art are also characterised by high levels of instruction.

Cummings (2010), for instance, conclude that field differences constitute the primary source of variance in faculty research publication. More specifically, publication rates in engineering, natural, and medical sciences are found to exceed those in social sciences and humanities. Piro, Asknes, and Rørstad (2013) confirm this pattern while also emphasising the effects of alternative counting methods. Once they employ fractional publication counts to account for higher numbers of co-authors in natural sciences, the picture clearly changes with humanities and social sciences ranking first and second, respectively. Field differences are even more pronounced when it comes to citation practices (Waltman, 2016). To illustrate this point, Radicchi, Fortunato, and Castellano (2008) state that receiving 100 citations is about 50 times more common in developmental biology than in aerospace engineering while Waltman et al. (2011) discover citation counts in biochemistry to be roughly one order of magnitude higher than in mathematics.

Moreover, there is evidence that educational processes are also subject to considerable heterogeneity. According to Smith and Naylor (2001a, 2001b), completion rates and degree results are both affected by field of study. For instance, receiving a good degree becomes more likely in humanities and biological sciences whereas dropout risk is increased in computer sciences, in each case relative to the baseline set by social sciences.[2]

In conclusion, disciplinary differences are substantial, take various forms, and have long been studied. Yet we are not aware of any study on higher education efficiency that has addressed this issue thoroughly. Some attempts were based on the distinct features of medical studies, which led to separating institutions with and without medical schools (Agasisti & Salerno, 2007; Ahn et al., 1988; Thanassoulis, Kortelainen, Johnes, & Johnes, 2011) or to adjusting data of medical schools (Hanke & Leopoldseder, 1998). In fact, Athanassopoulos and Shale (1997) might have presented the most comprehensive approach by dividing UK universities into three groups of different science orientation levels despite a rather arbitrary threshold setting. Overall, it seems that the ensuring homogeneity has mostly been overlooked.

While efficiency scores may thus be suffering from a considerable bias, subject mix has attracted (newfound) interest when it comes to explaining efficiency scores within two-stage frameworks. From an economics perspective, running a second-stage regression is of particular importance to gain insights into efficiency drivers on which grounds policy implications can be drawn. Medical faculties have frequently been included within these regressions (Agasisti & Pohl, 2012; Agasisti & Wolszczak-Derlacz, 2016; Kempkes & Pohl, 2010; Wolszczak-Derlacz, 2017; Wolszczak-Derlacz & Parteka, 2011) and in several cases found to have a significant impact. However, it may be questioned whether

[2] Both studies by Smith and Naylor include gender specific estimations and cover a variety of subjects. Their approach leads to numerous findings, which cannot be covered in detail at this point. The results presented are therefore rather illustrative albeit significant and representative for both genders.

universities can be commonly assessed without accounting for subject composition. Universities with a strong life sciences profile might simply be incapable of reaching an efficiency frontier composed of universities primarily engaged in social sciences. In fact, they might not even regard these universities as their peers, which essentially casts doubt upon the managerial side of relative efficiency techniques. Moreover, regression results become prone to misinterpretation once biased efficiency scores are utilised. Rather than being a source of inefficiency, medical faculties might be more likely to illustrate systematic differentials between academic fields that, of course, become more or less pronounced depending on the respective input-output specification.

3 Clustering Analysis

3.1 Methodology

There is a great number of clustering methods, from which we select the K-means algorithm. It is widely considered an elegant method for splitting a dataset into distinct clusters.[3] The idea behind K-means can be formalised in an intuitive way: Let $C_1, \dots, C_K$ denote K sets of distinct, non-overlapping clusters. Since clustering aims at grouping observations that tend to be similar, one can assess clusters based on their within-cluster variation, which should be as small as possible. The problem to be solved by the K-means algorithm can thus be stated as

$$\min_{C_1,\dots,C_K} \sum_{k=1}^{K} W(C_k), \tag{1}$$

where $W(C_k)$ denotes the within-cluster variation of cluster C_k. A common way to measure the within-cluster variation for C_k refers to the sum of squared distances between each observation $x \in C_k$ and the cluster's mean μ_k. Using squared Euclidean distance, we can redefine the optimisation problem as follows

$$\min_{C_1,\dots,C_K} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2. \tag{2}$$

While the logic underlying K-means clustering gives little cause of concern, one regularly faces the practical issue of selecting the parameter K. Since there is no well-accepted approach for this problem, we mainly follow Makles (2012) and determine the optimal number of clusters based on the total within-cluster variation, i.e. the target value of the

[3] Albeit the growing popularity of clustering applications in numerous fields, the higher education landscape has also only partly been explored by these techniques. Notable examples include Stanley and Reynolds (1994) and Valadkani and Worthington (2006) who evaluated performance differences within the Australian university system. In a similar vein, Shin (2009) grouped South Korean universities based on research performance while Bonaccorsi and Daraio (2009) and Lepori, Baschung, and Probst (2010) developed classification schemes for European universities.

optimisation problem above. More specifically, we consider any increase in K desirable as long as it is accompanied by a sufficient reduction of that value. For this purpose, we emphasise comparing the proportional reduction of error for different values of K. Formally, this coefficient is defined as

$$PRE(K) = \frac{WSS(K-1) - WSS(K)}{WSS(K-1)}, \tag{3}$$

where $WSS(K)$ denotes the total within-cluster variation for a solution of $K > 1$ clusters. Once this coefficient drops considerably, we refrain from partitioning our dataset any further. Additional explanations regarding our clustering methodology are partly given within the next subsection.

3.2 Data and Results

Our analysis covers the period from 2011 to 2014 and exploits two main data sources. The core data were derived from the ETER, which provides comparable micro-data on higher education institutions across Europe. In addition, we utilise data from Scopus, an abstract and citation database hosted by Elsevier, to supplement our institutional data with meaningful measures of research output. After restricting our dataset to public and government-dependent universities and eliminating specialist institutions (e.g. music and arts academies), a total of 450 universities from 16 European countries remains to constitute our sample.[4]

For our clustering analysis, we rely on publication records collected from Scopus. In principle, one could also argue in favour of employing student enrolment data for this purpose. Yet we consider research output the preferred choice primarily because our subsequent efficiency analysis addresses research activities in greater detail. Scopus does not only cover a broad range of scientific literature but also classifies its content under four main subject areas, i.e. life sciences, social sciences, physical sciences, and health sciences. Building on this classification system, we determine a university's share of publications in each of these four subject areas, which constitutes its position in subject space. These vectors then serve as the foundation for our clustering analysis. However, it should be noted that Scopus' subject areas are partly overlapping. For instance, publications in *Applied Mathematical Finance* are assigned to both social and physical sciences. Multidisciplinary work therefore entails the potential risk of distorting subject profiles towards research areas with greater overlap. For this reason, we employ a fractional counting approach, which essentially leads to splitting contributions equally between subject areas. On average, universities in our sample account for research output

[4] ETER classifies an institution as government-dependent if either more than 50% of its core funding is provided by government agencies or its staff is paid by a government agency (Lepori et al., 2016, p. 38).

15% of which fall under life sciences, 19% under social sciences, 46% under physical sciences, and 20% under health sciences.[5]

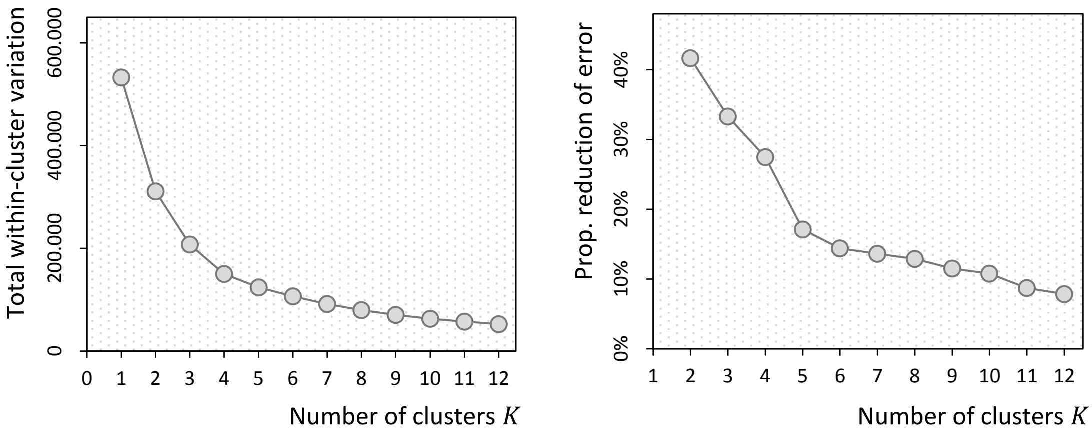

Fig. 1: Total within-cluster variation (left) and proportional reduction of error (right)
Notes: Values are averaged over 1 000 replications of K-means with random starting centres.

The task of selecting an appropriate number of clusters is addressed in Figure 1. The left-hand panel depicts the total within-cluster variation for different values of $K.$ Raising the number of clusters apparently reduces variation; however, there is a diminishing benefit along with it. This effect becomes even more evident in view of the right-hand panel of Figure 1 that displays the proportional reduction of error. Based on this criterion, adding a second cluster has the biggest impact reducing total within-cluster variation by 42%. By adding a third and fourth cluster, variation continues to drop by 33% and 28%, respectively. Afterwards, the graph shows a rather steep decline. An additional fifth cluster would decrease variation by (merely) 17%. Moreover, the impact would barely differ from adding a sixth, seventh, or eighth cluster, so that any of these solutions would appear rather arbitrary. Although deciding on the optimal number of K is usually not a clear cut, the statistics are mostly supportive of a four-cluster solution in our case. This also includes the Calinski–Harabasz (1974) pseudo-F index that we calculated as an additional check (see Appendix A for details).

The final step of our clustering approach is to apply the actual K-means algorithm. One shortcoming of K-means arises from the fact that the algorithm converges to local instead of global optima. We therefore ran the algorithm multiple times with random starting centres and selected the solution with minimal objective value as suggested by

[5] Physical sciences appear to be overrepresented, which is partly attributable to database coverage. However, this bias is found to be even larger within the Web of Science, which may have served as an alternative data source. Besides, broad-scale comparison reveals that Scopus exceeds Web of Science in terms of journal coverage in every disciplinary field (see Mongeon and Paul-Hus, 2016, on both aspects) and thus provides a more reliable basis for accurate efficiency estimations.

James, Witten, Hastie, and Tibshirani (2013, pp. 388–389). The final clustering was obtained in 51 out of 1 000 replications. Aggregate statistics on subject space location by cluster are presented in Table 1.

Cluster	***N***	**Life Sciences**	**Social Sciences**	**Physical Sciences**	**Health Sciences**	**Focus**
CLUSTER 1	57	6.25%	56.10%	22.21%	15.11%	*Social Sciences*
CLUSTER 2	140	8.49%	10.74%	74.02%	6.30%	*Physical Sciences*
CLUSTER 3	49	21.41%	12.41%	14.43%	51.24%	*Health Sciences*
CLUSTER 4	204	20.15%	14.95%	40.80%	23.32%	*General*
Sample	450	14.90%	18.57%	45.91%	20.02%	-

Tab. 1: Mean composition of research output by cluster

According to Table 1, the European public university landscape can hardly be regarded as homogeneous. In fact, there are significant differences in terms of subject focus. This becomes particularly evident with regard to specialist clusters such as CLUSTER 1. On average, 56% of the publications by a CLUSTER 1 university belong to social sciences. In contrast, other clusters display mean values that are up to five times smaller ranging from 11 to 15%.[6] A similar degree of specialisation can be observed by CLUSTER 3, which comprises universities that lay its emphasis on health sciences. These two clusters also resemble each other in terms of size, as they are considerably smaller than the remaining clusters. Albeit covering a lot more universities, CLUSTER 3 can still be viewed as a specialist cluster primarily directed towards physical sciences. Lastly, CLUSTER 4 contains universities that most closely resemble the sample mean (thereby sharing the general tendency towards physical sciences). We thus regard them as generalist institutions.

[6] It should be noted that mean values might be misleading to some extent. For instance, a few universities outside of CLUSTER 1 are visibly engaged in social sciences along with their primary cluster focus. Yet only two of them marginally exceed the lower bound of CLUSTER 1 set by the Birmingham City University (35.57%). Still, this observation could be indicative for cluster boundaries being partly fluid (see Chapter 7).

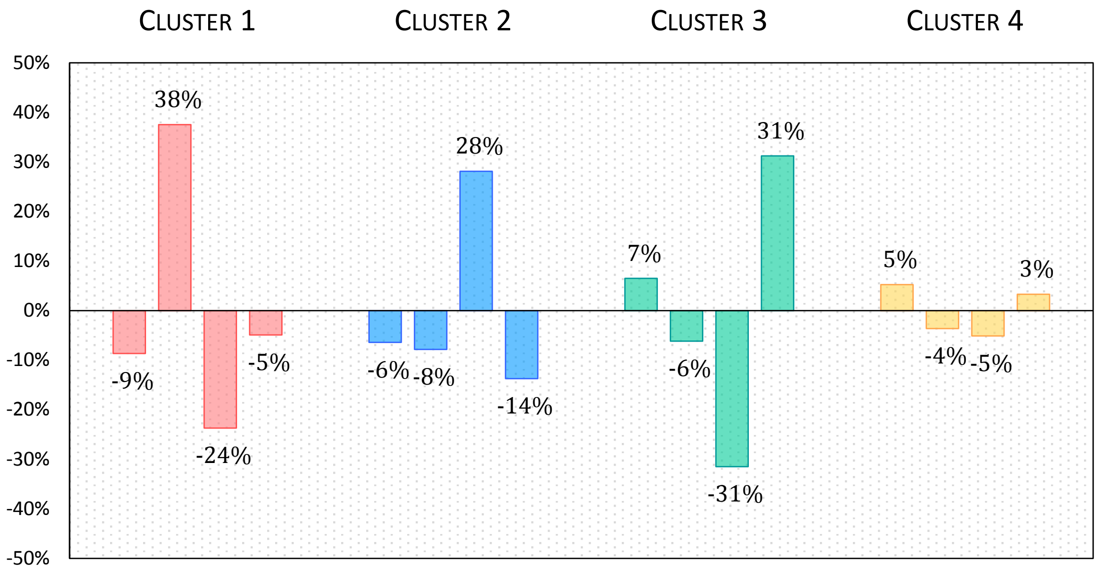

Fig. 2: Subject deviation by cluster in comparison to sample mean
Notes: Bar order = life sciences, social sciences, physical sciences, health sciences.

Cluster characteristics are further illustrated by Figure 2, which depicts how far clusters deviate from the sample mean, and by Figure 3, which emphasises comparisons within the subject space boundaries effectively set by our data. The latter approach is particularly relevant given that the maximum degree of specialisation notably differs between subject areas. For instance, while we discover universities reaching output fractions of above 90% in social and physical sciences, peak values in life and health sciences lie within the 60% and 70% region, respectively. Employing an identical scale along each subject dimension might therefore conceal insights. Instead, we used linear transformations to map our data onto the intervals ranging between the 1st and 99th percentile.[7] Following this approach, it first becomes clear that our initial result holds true: Three clusters show a distinct subject focus. Yet differences in the degree of specialisation appear in a different light. For instance, Figure 3 reveals a comparable level of specialisation for CLUSTER 2 and 3, which is primarily due to higher expansion on the health sciences axis. In other words, both clusters become increasingly similar if we acknowledge that specialisation in the field of health sciences relates to lower output fractions than in physical sciences. This effect is indeed most pronounced in life sciences where values above 50% rarely occur. As a result, universities in CLUSTER 4 appear increasingly balanced hence approaching the perception of generalist institutions more closely.

[7] Standard rescaling refers to utilising minimum and maximum values. As a result, this approach is known to be sensitive to outliers. Percentile ranks thus provide a reasonable alternative.

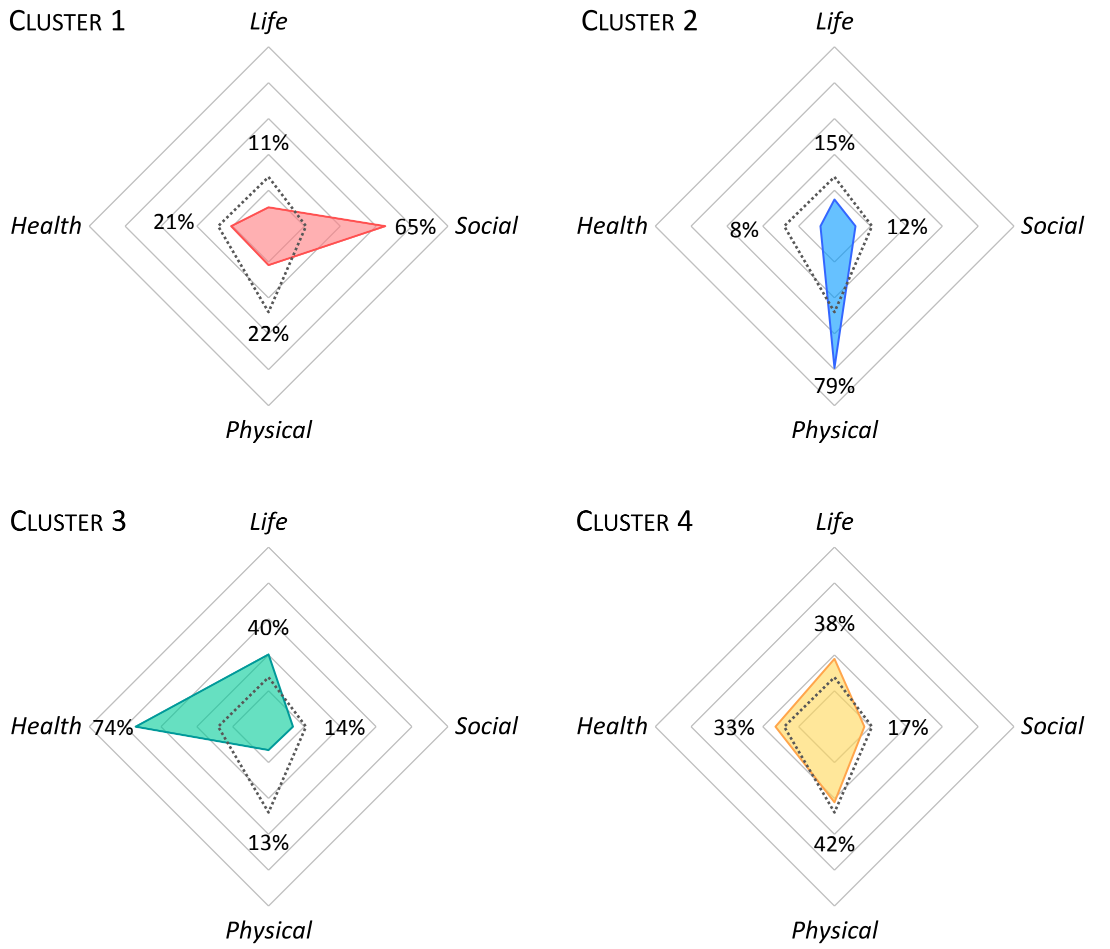

Fig. 3: Subject focus by cluster in comparison to sample mean (dotted line)
Notes: Data are rescaled to lie between the 1st and 99th percentile, axes range from 0-100%.

The clustering analysis clearly sheds light on systematic differences between groups of universities in Europe. It is interesting to note that we hereby partly confirm the results by Lepori et al. (2010), who identify specialised institutions in the fields of technical-natural sciences and social sciences and humanities. Differences in subject space alone could indeed be overlooked by efficiency analyses if they were not linked to further institutional disparities. However, descriptive statistics presented in Table 2 point to the contrary. Referring to the output dimension, publications per academic staff, measured as full-time equivalents, constitute a typical indicator for scientific productivity. In line with the cited literature, we discover the lowest values within the social sciences cluster. On average, we find these universities to record an annual number of 0.49 publications per academic employee between 2011 and 2014.[8] In comparison, mean values of 0.83 and 0.91 are achieved by universities focussing on physical and health sciences, which

[8] Our study is not restricted to research articles but includes every publication format on Scopus. This is particularly relevant for the field of social sciences and humanities where books and book chapters are known to play an important role in scientific communication.

represents an increase of 69% and 86%, respectively. Overall, the general cluster is associated with the highest productivity of 1.01 publications per academic staff, which may point to the existence of economies of scope. A similar picture emerges with regard to the number of citations per publication, which we include to capture the impact of scientific contributions. Based on an evaluation window that covers the publication year plus two subsequent years, we find an average citation rate of 3.13 within the social sciences cluster, which amounts to less than half of what their counterparts with a health sciences or general profile are able to accomplish.

With respect to the input dimension, we direct attention to current expenditures as a summary measure of resource usage. By comparing the annual expenditure levels per student, we clearly observe the health sciences cluster to be a costly exception. In contrast, universities from the social sciences cluster record relatively low expenditure numbers despite being exposed to a higher teaching load (as indicated by the ratio of students to academic staff). Again, both of these findings are consistent with the reviewed literature. Lastly, we see that universities focussing on social and health sciences are of similar size accommodating an average of 11 to 12 thousand students. In comparison, we find the number of students to exceed 15 thousand in the physical sciences cluster and approach 22 thousand among generalist universities.

It is crucial to note that systematic differences between clusters are a major cause for concern from the standpoint of efficiency analysis. More specifically, they suggest that production processes are subject to heterogeneous technologies, which would remain unconsidered if universities were pooled together across the entire subject spectrum. Instead, we strongly argue in favour of performing efficiency estimation cluster-wise to ensure a comparison of (true) peers.

Cluster	Publications per academic staff	Citations per publication	Students per academic staff	Expenditures per student	Number of students
1 – Social Sciences					
P5	0.06	1.32	11.09	3 569	1 343
Mean	0.49	3.13	20.91	9 911	11 029
P95	1.35	5.91	32.00	20 199	22 945
2 – Physical Sciences					
P5	0.17	1.91	5.41	3 435	3 043
Mean	0.83	4.86	16.89	12 564	15 575
P95	1.77	9.44	29.12	31 530	35 798
3 – Health Sciences					
P5	0.12	1.87	1.47	4 341	1 608
Mean	0.91	6.73	15.69	38 836	11 921
P95	2.68	12.39	30.84	213 706	30 027
4 – General					
P5	0.32	2.77	5.17	5 228	6 928
Mean	1.01	6.81	15.13	15 189	21 882
P95	1.82	11.01	28.33	38 542	48 150
Sample					
P5	0.16	1.94	5.15	3 944	3 059
Mean	0.88	5.73	16.47	8 933	17 461
P95	1.81	10.72	29.17	39 265	38 515

Tab. 2: Summary statistics by cluster

Notes: Publications comprise all document types listed on Scopus. The citation window covers three years including the year of publication. Academic staff is expressed in FTE. Financial data are converted into real PPP EUR (2014 = 100).

4 Statistical DEA Approach

We employ a statistical DEA approach in line with Simar and Wilson (1998, 2000). Thus, let $x \in R_{+}^{p}$ denote a vector of p inputs and $y \in R_{+}^{q}$ a vector of q outputs. The production possibilities set can then be defined as

$$P = \{\, (x, y) \in R_{+}^{p} \times R_{+}^{q} \mid x \; can \; produce \; y \,\}. \tag{4}$$

Production facilities, in our case universities, in the interior of P are termed technically inefficient, whereas universities located on the boundary, or frontier, of P are considered technically efficient. In order to determine the degree of efficiency, we adopt an output-oriented perspective implicitly assuming that universities have greater control over outputs than inputs.[9] A university located at a given point (x, y) can thus be assessed by

[9] This view is shared by a number of studies, including those by Agasisti and Johnes (2009), Kempkes and Pohl (2010), and Wolszczak-Derlacz and Parteka (2011).

$$\theta(x, y \mid P) = sup\,\{\, \theta > 0 \mid (x, \theta y)\, \epsilon\, P \,\}, \tag{5}$$

where $\theta(x, y \mid P)\ \epsilon\ [1, \infty)$ measures the largest radial expansion of y that is feasible given x. Higher inefficiency is accordingly indicated by larger values of $\theta(x, y \mid P)$. In theory, inefficiency scores could be obtained through mathematical programming if the set of production possibilities were fully disclosed. However, this is not the case. Instead of observing all possible input-output combinations, one generally encounters a subset of technologies from P, denoted by $\hat{P}$. We thus refer to P and $\theta(x, y \mid P)$ as the true but unknown quantities of interest and to $\hat{P}$ and $\theta\left(x, y \,\middle|\, \hat{P}\right)$ as their sample estimators.

By construction, $\hat{P}$ constitutes an inner approximation of P, which causes inefficiency estimates to be downward biased, i.e. $\theta\left(x, y \,\middle|\, \hat{P}\right) \leq \theta(x, y \mid P)$. In dealing with this issue, one generally relies on bootstrap-based inference. This leads to a virtual environment, where $\hat{P}$ and $\theta\left(x, y \,\middle|\, \hat{P}\right)$ become the quantities of interest to be estimated by P^* and $\theta(x, y \mid P^*)$, which build on subsets drawn from the original data. Further, let $\hat{F}$ refer to the bootstrap data generating process that mimics the true data generating process F. It then follows that

$$\theta\left(x, y \,\middle|\, \hat{P}\right) - \theta(x, y \mid P^*) \,\middle|\, \hat{F} \;\sim\; \theta(x, y \mid P) - \theta\left(x, y \,\middle|\, \hat{P}\right) \,\middle|\, F, \tag{6}$$

so that a bias-corrected estimator of $\theta(x, y \mid P)$ can be stated as

$$\tilde{\theta}(x, y) = 2\, \theta\left(x, y \,\middle|\, \hat{P}\right) - E\big(\theta(x, y \mid P^*)\big). \tag{7}$$

Technically, we employ the homogeneous bootstrap algorithm proposed by Simar and Wilson (1998) based on 1 000 replications. In addition, we assume free disposability along with convexity and allow for variable returns to scale when constructing estimates of P. This procedure leads to bias-corrected efficiency scores $\{\tilde{\theta}_i : i = 1, \ldots, n\}$ for our set of n universities, which we examine descriptively and further by means of kernel densities. For this reason, let f denote the density of $\tilde{\theta}$. Its standard kernel density estimator at any point u is then defined as

$$\hat{f}(u) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{u - \tilde{\theta}_i}{h}\right), \tag{8}$$

where $K(\cdot)$ denotes a kernel function and h constitutes a suitable bandwidth. However, due to efficiency scores being constructed with bounded support, this estimator requires alteration to ensure consistency. We therefore apply the modified estimator

$$\hat{f}_R(u) = \begin{cases} \dfrac{1}{nh_R} \displaystyle\sum_{i=1}^{n} \left[K\left(\dfrac{u - \tilde{\theta}_i}{h_R}\right) + K\left(\dfrac{u - (2 - \tilde{\theta}_i)}{h_R}\right)\right], & u \geq 1 \\ 0, & otherwise, \end{cases} \tag{9}$$

where h_R denotes an adjusted bandwidth. Moreover, we opt for a Gaussian kernel and follow Silverman's rule (1986) for bandwidth selection.[10]

In a second stage, we investigate potential efficiency drivers by employing the bootstrap regression framework by Simar and Wilson (2007). More precisely, we expect university i's true efficiency θ_i to depend on a vector z_i of covariates. On the assumption that these covariates exert constant percentage effects, our model resolves to

$$ln(\theta_i) = \psi\,(z_i, \beta) + \varepsilon_i, \tag{10}$$

where β denotes a vector of coefficients, $\psi\,(\cdot)$ describes a functional form later to be determined, and ε_i represents the unexplained residual term, which is assumed to be normally distributed with left-truncation at $-\,\psi\,(z_i, \beta)$. We estimate this model by means of maximum likelihood, again based on 1 000 replications. With true efficiencies remaining unknown, we rely on their bias-corrected estimates for inference about β.

5 Estimating Efficiency

5.1 Model Specification

Universities are generally known to engage in two major fields of activity, i.e. teaching and research. With regard to teaching, we include the number of yearly graduates on ISCED levels 5 to 7 as our preferred output measure. Some studies instead opted for the number of enrolled students noting that education received by students who drop out before graduation should not be neglected (Cohn, Rhine, & Santos, 1989). However, this approach could be prone to misjudgement caused by inactive students (so-called phantom students), which are of particular relevance when evaluating public institutions (Teixeira, Rocha, Biscaia, & Cardoso, 2012). In fact, universities that handle these students least effectively hereby revealing inefficient administrations were the primary beneficiaries of this measure. Besides, study efforts that fail to be rewarded with degrees should not be overstated since job markets returns are considerably decreased after the event of dropping out (Walker & Zhu, 2013).

To capture research activities, we include the number of scientific publications in our model. They are indeed central for knowledge dissemination as scientific contributions usually manifest in some form of publication. However, it seems reasonable to argue that publications only serve as a partial indicator for research output. Following Martin and Irvine (1983), we regard them mainly as a measure of scientific production but not of scientific progress. This distinction rests upon the notion that publications tend to vary in scientific value. Most contributions might be incremental in nature, while some add considerably to the advancement of science. Individuals but also institutions as a

[10] See Sickles and Zelenyuk (2019) for further details on density estimation in efficiency contexts.

whole might have different preferences and abilities regarding these two dimensions, which calls for an additional output measure. We therefore incorporate citation counts hereby broadening the scope of previous efficiency studies that usually omitted this measure due to data limitations.[11] While it might be tempting to consider citations a measure of quality, we again support the scientometric view and regard them rather reflective of scientific impact (Martin & Irvine, 1983). Furthermore, one might speak of short-term impact given that our citation window is restricted to a maximum of three years. Moed et al. (1985) point out that not every contribution to the current research front eventually becomes accepted knowledge, which motivates the distinction between short-term and long-term impact. However, there is empirical evidence that suggests that both concepts are closely linked to each other (Adams, 2005). We further clarify this relation by providing a separate analysis that underlines the significant correlation between initial and overall citations at the institutional level (see Appendix B). Based on these results, our efficiency estimates are expected to be robust to extended citation windows.

In contrast to the output side, the literature reveals less consistency over the choice of inputs. This becomes especially apparent by the variety of expenditure types that have been utilised, e.g. including expenditure on personnel, central administration, or library services. These differences may partly result from the availability of data but, more importantly, express alternative views of higher education efficiency. Our take on this matter is rather strict. In line with Thanassoulis et al. (2011), we define the level of current expenditures (converted in purchasing power parities) as our single input. Two main reasons can be pointed out in support of this approach. First, from a public finance standpoint, it seems hardly relevant in what specific way resources are allocated within an institution. Universities are generally given a great amount of (operational) freedom, which can shape production processes in various forms, with labour or capital intensive organisation serving as classic examples. However, exploring how efficient universities are making use of certain resources is at best of secondary concern for policy makers. Their focus is expected to lie on the overall budget. Second, there are technical reasons for limiting the number of inputs. DEA is a flexible technique that allows units to attach individual weights to input and output components so that they appear in the most favourable light relative to their peers. Broadening the set of inputs would therefore open up more opportunities for universities to become efficient, which we consider unreasonable. To illustrate this point, theoretically adding the number of students as a

[11] Citation counts have frequently been advocated but rarely been included to capture research output. To the best of our knowledge, only one study that exploited citation data for efficiency purposes, i.e. Bonaccorsi, Daraio, and Simar (2006), who examined the Italian university system.

second input dimension would permit universities to be assessed based on their citations to students ratio, which would be at odds with our efficiency perception.

Overall, our model includes publications, citations, and graduates as outputs and current expenditures as an aggregate input measure. Moreover, we employ a fractional counting approach meaning that credit for publications and citations is split between collaborating universities according to the number of contributing authors. In view of a potential bias related to varying centrality within our European university network, it is worth noting that we consider co-authorship ties to any other affiliation for this task and not only links to universities from our sample.

5.2 Results

This sections proceeds with exploring the results of our efficiency estimation that we conduct yearly between 2011 and 2014. While we strongly advocate applying cluster-specific technology frontiers, we also present results based on a global frontier to contrast both approaches.

Cluster	***N***	**P5**	**P25**	**P50**	**Mean**	**P75**	**P95**	**SD**
Cluster-specific Frontiers								
Social	228	1.15	1.30	1.64	2.04	2.45	4.23	1.11
Physical	560	1.16	1.33	1.75	2.05	2.47	4.03	1.02
Health	196	1.14	1.32	1.67	2.22	2.41	5.03	1.47
General	816	1.18	1.47	1.81	2.03	2.27	3.75	0.81
Global Frontier								
Social	228	1.16	1.71	2.20	2.60	2.97	5.39	1.52
Physical	560	1.16	1.41	1.96	2.23	2.66	4.31	1.10
Health	196	1.22	1.75	2.32	3.00	3.53	7.35	2.14
General	816	1.18	1.53	1.90	2.16	2.43	4.15	0.91

Tab. 3: Summary statistics on bias-corrected efficiency scores by cluster

Table 3 provides summary statistics on efficiency estimates aggregated by cluster and pooled over our 4-year observation period. From the upper half of this table, we can infer that assessing universities cluster-wise reveals efficiency distributions with a high degree of similarity (with one minor exception being the health cluster where the low-performance segment, i.e. the long tail, appears slightly more accentuated). In contrast, results derived from a global frontier lead to conclude that efficiency differs notably between clusters. This finding becomes increasingly visible as percentile ranks increase. There are indeed highly efficient universities within each cluster; however, beyond the 5th percentile, we see an efficiency gap widening between the social and health cluster on the one hand and the physical and general cluster on the other hand. In order to evaluate this gap in more detail, we employ the non-parametric Kruskal-Wallis test, which clearly rejects the null hypothesis of equal mean ranks across clusters (χ^2 = 46.986 with an associated p-value of 0.0001). In light of this significant result, one might draw the conclusion that some clusters simply outperform others. Again, we consider this rather a display of unreasonable comparison.

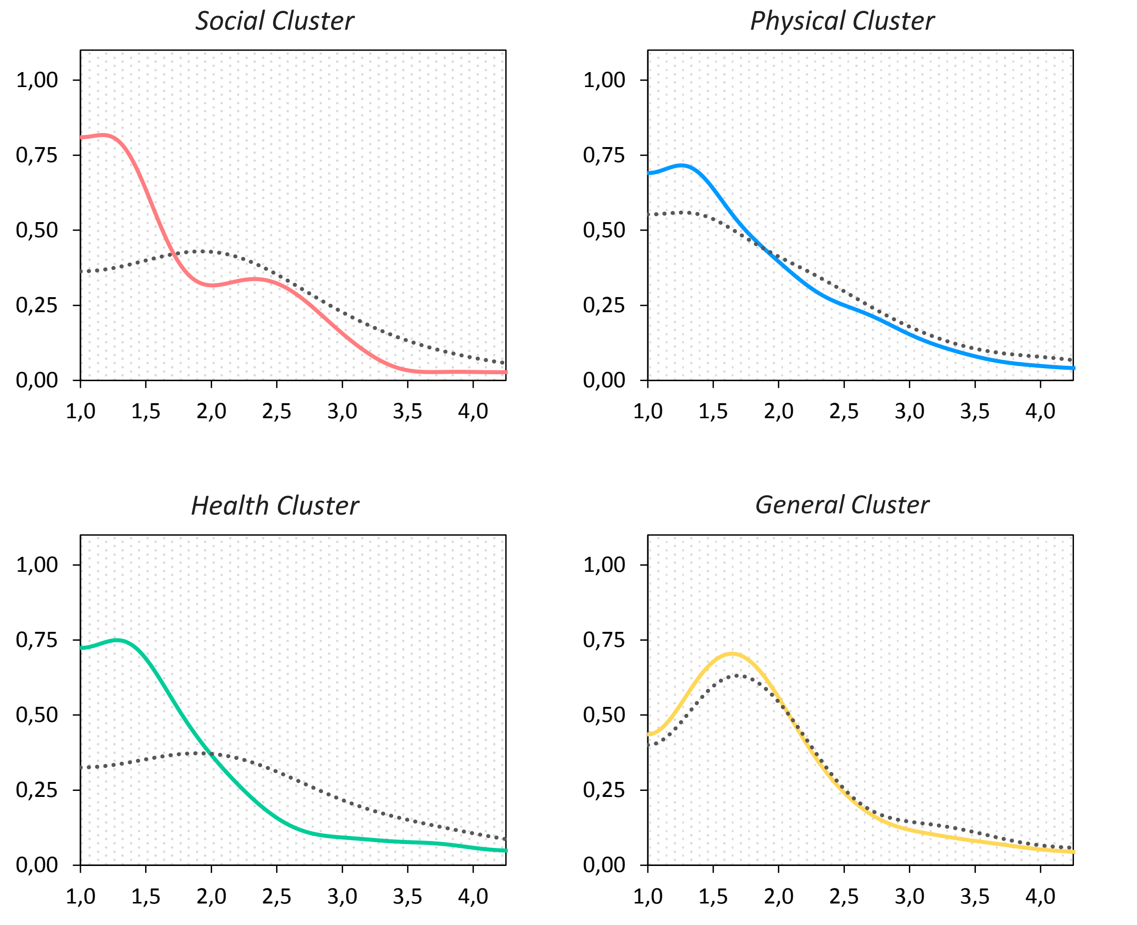

Fig. 4: Density estimates of bias-corrected efficiency scores by cluster

Notes: Densities refer to estimates derived from an individual cluster frontier (solid line) or from one global frontier (dotted line).

In addition to Table 4, density estimates of bias-corrected efficiency scores are visualised in Figure 4. Three observations are worth emphasising here. First, and in line with our previous findings, switching from a global to an intra-cluster frontier impacts universities focused on social and health sciences the most. In the latter scenario, more probability mass becomes assigned towards unity, which is (to a certain extent) also due to the relatively high reduction in sample size. Second, all distributions appear to be right-skewed, which marks a frequently expected outcome in efficiency contexts, and leptokurtic. Third, we observe a wide range of efficiency estimates including some extreme values, which indicates considerable heterogeneity among universities but also high discriminatory power of our model.

Country	**Global Frontier**		**Social Cluster**		**Physical Cluster**		**Health Cluster**		**General Cluster**	
	Mean	P50	Mean	P50	Mean	P50	Mean	P50	Mean	P50
Belgium	1.71	1.57	-	-	-	-	-	-	1.47	1.43
Czech Republic	2.10	2.00	-	-	1.93	1.79	-	-	1.74	1.79
Finland	2.45	2.32	2.09	2.08	1.96	1.76	-	-	2.28	2.23
Germany	3.27	3.12	2.48	2.52	2.72	2.55	3.67	3.74	3.08	3.09
Great Britain	1.92	1.79	1.78	1.49	1.84	1.76	1.44	1.33	1.72	1.68
Ireland	1.80	1.80	-	-	-	-	-	-	1.71	1.63
Italy	2.09	1.76	4.22	3.64	1.95	1.69	1.47	1.39	1.71	1.62
Lithuania	5.64	4.55	-	-	3.79	3.64	-	-	-	-
Netherlands	1.77	1.77	-	-	1.34	1.25	-	-	1.61	1.76
Norway	2.65	2.59	-	-	-	-	-	-	2.39	2.31
Poland	1.50	1.39	1.21	1.18	1.43	1.33	1.45	1.41	1.40	1.35
Portugal	1.91	1.81	-	-	1.57	1.20	-	-	1.72	1.79
Sweden	2.96	2.87	-	-	2.38	2.21	2.11	2.05	2.42	2.20
Switzerland	2.82	2.36	-	-	1.98	1.35	-	-	2.48	2.20
Sample	2.33	1.99	2.04	1.64	2.05	1.75	2.22	1.67	2.03	1.81

Tab. 4: Mean and median bias-corrected efficiency scores by country and cluster
Notes: Efficiencies scores referring to less than three institutions are not reported. Malta and Cyprus are left out for this reason.

Exploring efficiency levels from a national perspective reveals further insights. According to Table 4, mean and median efficiency scores show substantial variation across Europe. The group of top-performing countries mainly comprises Poland, Belgium, and the Netherlands. Relatively high efficiency scores are also realised by universities in Portugal, Czech Republic, Great Britain, and Ireland, whereas Scandinavian universities generally offer more room for improvement. Apart from these general patterns, there are cluster deviations worth emphasising. Italy, for instance, achieves high efficiency in the health sciences cluster, while clearly lagging behind in the social sciences cluster. Interestingly, the reverse picture emerges with regard to Germany despite its overall greater level of inefficiency. Lastly, Switzerland appears more efficient in the physical sciences cluster than in the general cluster.

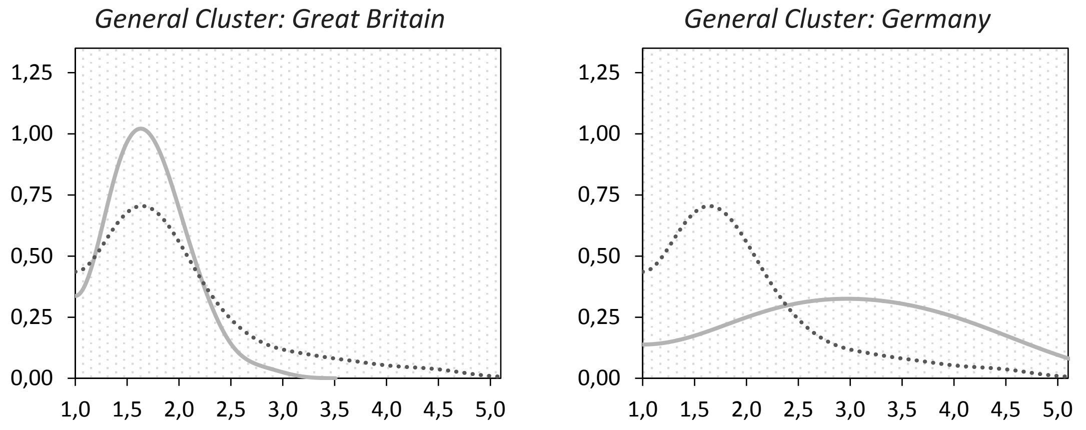

Fig. 5: Density estimates of bias-corrected efficiency scores by country, Cluster = General

Notes: Densities refer to a single country (solid line) or to the overall cluster (dotted line).

To provide an outlook beyond measures of central tendency, we recommend taking a closer look at the general cluster and, more specifically, at universities from Germany and Great Britain, which represent its largest subgroups. Based on density estimates shown in Figure 5, both countries can be considered to differ not only in mean or median efficiency but also in terms of within-country variation. Clearly, the German university landscape reveals a lot more heterogeneity than its British counterpart does. Upon examining the remaining countries, it seems difficult to state a general rule. Yet the illustrated examples seem to be representative in the sense that high mean inefficiency usually indicates more variation. Additional density plots for countries with meaningful sample sizes are provided in Appendix C.

6 Exploring Efficiency Drivers

6.1 Model Specification

From a policy perspective, detecting inefficiencies in public institutions can only be seen as an intermediate step. The focus of this section will therefore be placed on identifying efficiency drivers, knowledge of which may be useful for designing reasonable policy measures to promote higher education efficiency. Overall, we consider the following model specification

$$\begin{aligned} ln(\tilde{\theta}_{it}) = \; & \beta_0 + \beta_1 ln(Size_{it}) + \beta_2 Herfindahl_{it} + \beta_3 Thirdparty_{it} \\ & + \beta_4 Fees_{it} + \gamma X'_{it} + \alpha_i + \delta_t + u_{it}, \end{aligned} \tag{11}$$

which relates university i's efficiency estimate in year t to various factors expected to be of influence. In particular, we are interested in the potential effects of university size approximated by the number of students ($Size$), subject specialisation calculated as a Herfindahl index ($Herfindahl$), and funding composition characterised by the share of current revenues raised through third-party funds ($Thirdparty$) and student fees ($Fees$). Moreover, we include a set of year dummies (δ) to control for time fixed effects, a set of country dummies (α) to account for country fixed effects, and further controls (X) related to employee structure, institutional design, and regional productivity. Summary statistics on all covariates by cluster along with more precise descriptions are provided in Table 5. Note that these data are derived from ETER except for GDP, which originates from Eurostat and the Swiss Federal Statistical Office.

Among our variables of interest, $Size$ permits investigating potential economies of scale in higher education. From a theoretical standpoint, large universities might benefit from higher utilisation of various assets. These could include shared research infrastructure, e.g. production plants or computing centres that typically require considerable initial investments, but also educational facilities such as libraries. Moreover, advancements in information technology could lead to a reduced demand of interpersonal relations in teaching hence expanding the range of decreasing unit costs presumably in favour of large institutions that tend to offer lectures for greater student numbers. However, administrative tasks potentially are a source of diseconomies of scale since organisational costs are expected to increase disproportionately with size. In view of these opposing arguments, it is understandable that the empirical literature has not yet reached a consensus on this matter (Bonaccorsi, Daraio, & Simar, 2006).

We further aim to shed light on economies of scope by including $Herfindahl$ as a measure of subject specialisation in our model. Based on ETER's distinction of 11 fields of study, this index ranges between 0.1, if students are equally distributed across fields, and 1.0, if students belong to exactly one field. Even though the Herfindahl index rests upon student numbers, it largely resembles our clustering results. As can be seen from Table 5, specialised clusters are characterised by higher index values hereby providing a first indication of the robustness of our approach. Whether efficiency benefits from specialisation or diversification in subject coverage is hard to answer on theoretical grounds. Turning to empirical studies, the overall picture remains mostly unclear. According to Daraio et al. (2015a), specialisation enhances academic efficiency whereas results from Agasisti and Wolszczak-Derlacz (2016) as well as Wolszczak-Derlacz (2017) point to the contrary, i.e. the presence of economies of scope. Yet another conclusion is derived by Bonaccorsi et al. (2006), who reject any significant relation.

Lastly, our interest lies in evaluating if differences in funding structure are related to university efficiency. Although external funding has become an increasingly central revenue source for European universities, empirical evidence on its performance impact remains relatively scant. Still, we expect universities with larger proportions of third-party funds to be more efficient given that previous studies by Wolszczak-Derlacz and

Variable	**Sample**	**Social Cluster**	**Physical Cluster**	**Health Cluster**	**General Cluster**
GDP –	*Regional gross domestic product per capita according to NUTS 2 classification*				
P5	16 005	18 802	14 048	17 180	17 180
Mean	32 257	46 018	26 887	38 988	30 480
P95	48 400	159 662	46 954	157 583	47 858
Multisite –	*Binary variable indicating campuses outside a university's main location*				
P5	0.00	0.00	0.00	0.00	0.00
Mean	0.24	0.20	0.25	0.19	0.26
P95	1.00	1.00	1.00	1.00	1.00
Hospital –	*Binary variable indicating the presence of a university hospital*				
P5	0.00	0.00	0.00	0.00	0.00
Mean	0.29	0.04	0.07	0.43	0.47
P95	1.00	0.00	1.00	1.00	1.00
Size –	*Number of enrolled students at ISCED levels 5-7*				
P5	3 059	1 343	3 043	1 608	6 928
Mean	17 461	11 029	15 575	11 921	21 882
P95	38 515	22 945	35 798	30 027	48 150
Herfindahl –	*Herfindahl index based on enrolled students at ISCED levels 5-7 by subject (in %)*				
P5	14.24	15.62	15.34	16.27	13.77
Mean	26.95	37.25	31.17	37.75	18.75
P95	75.27	97.58	76.55	99.80	27.51
Prof –	*Proportion of full professors amongst employees (in %)*				
P5	2.22	1.39	3.03	1.74	2.42
Mean	6.54	6.09	6.91	6.42	6.49
P95	11.32	10.34	11.44	12.12	11.40
Female –	*Proportion of women amongst full professors (in %)*				
P5	7.69	9.60	4.92	6.67	12.29
Mean	22.50	31.54	17.39	24.90	22.80
P95	40.00	50.00	30.38	50.00	36.36
International –	*Proportion of foreigners amongst academic employees (in %)*				
P5	1.74	4.97	1.11	1.04	1.86
Mean	16.17	19.54	14.72	14.15	16.37
P95	41.16	55.70	45.87	39.60	39.72
Thirdparty –	*Proportion of current revenues raised through third-party funds (in %)*				
P5	1.32	0.66	1.16	1.87	2.21
Mean	17.47	11.65	21.16	20.90	16.43
P95	40.34	36.69	42.32	50.06	35.83
Fees –	*Proportion of current revenues raised through student fees (in %)*				
P5	0.13	0.58	0.12	0.05	0.14
Mean	23.02	41.39	14.52	16.90	23.57
P95	69.36	76.05	49.05	68.50	67.71

Tab. 5: Description and summary statistics on covariates by cluster

Notes: Financial data are converted into real PPP EUR (2014 = 100). Breakdown of employee structure is based on headcounts.

Parteka (2011), Agasisti and Wolszczak-Derlacz (2016), and Wolszczak-Derlacz (2017) discovered a negative correlation between the share of core funding and university performance in cross-country contexts. We further extend this strand of research by including the share of student fees, which allows us to disentangle the overall effect of external funding into two separate components. Although employing a parametric model, Bolli et al. (2016) pursue a similar approach and conclude that different mechanisms are potentially in play for these funding sources. More precisely, the share of tuition fees is found to decrease university efficiency while the opposite is revealed about international public funds.

Apart from investigating a comprehensive set of efficiency drivers, our methodological framework is in particular designed to uncover differences between subject areas. As indicated by our clustering analysis, universities likely operate under varying technological constraints, which casts doubt on assuming that covariates exert identical effects across fields. For instance, multidisciplinary work could be of different value across domains. Instead of jointly testing for economies of scope, we thus recommend evaluating clusters on an individual basis.

6.2 Results

The results of the regression analysis are reported in Table 6. In line with the previous chapter, our focus is twofold. We present the results of our preferred approach that builds on cluster-segmented estimation but also contrast it with the outcome of the pooling approach, which derives efficiency estimates from a global technology frontier. It should hereby kept in mind that our dependent variable constitutes a measure of inefficiency rather than efficiency. Coefficient estimates with a negative sign therefore indicate efficiency-enhancing effects while a positive sign corresponds to efficiency-decreasing effects.

Our first result is indeed not linked to a single variable but associated with the overall effect heterogeneity, which we find to take various forms. For instance, specialisation supposedly increases university efficiency in the pooled model as indicated by the significant and negative coefficient of $Herfindahl$. However, the segmented approach solely confirms this relation in case of the physical cluster. In a similar vein, higher shares of foreigners amongst academic employees ($International$) are associated with lower efficiency in the pooled model. Not only does this notion appear too general in view of the segmented analysis, it might potentially be misleading. While the effect points to the same direction within the social cluster, universities in the health cluster seem to benefit from increasing levels of internationalisation. Furthermore, we observe the general cluster to be an exception regarding the effects of $Female$ in the sense

Variable	**Global Frontier**	**Social Cluster**	**Physical Cluster**	**Health Cluster**	**General Cluster**
Natural logarithm of bias-corrected efficiency score as dependent variable					
ln(GDP)	- 0.0412	- 0.1500 **	0.1151	0.0923	- 0.0735 *
	(0.0237)	(0.0466)	(0.0706)	(0.0746)	(0.0362)
Multisite	0.0275	0.0646	0.1108 *	0.1947	0.0095
	(0.0203)	(0.0534)	(0.0484)	(0.1091)	(0.0208)
Hospital	0.0867 ***	- 0.2130	- 0.0190	0.2849 ***	0.1065 ***
	(0.0197)	(0.1302)	(0.0624)	(0.0646)	(0.0233)
ln(Size)	- 0.2843 ***	- 0.2909 ***	- 0.3985 ***	- 0.1034 *	- 0.1737 ***
	(0.0127)	(0.0469)	(0.0268)	(0.0484)	(0.0199)
Herfindahl	- 0.0029 ***	0.0018	- 0.0033 **	- 0.0026	- 0.0012
	(0.0006)	(0.0020)	(0.0010)	(0.0016)	(0.0029)
Prof	- 0.0212 ***	- 0.0255 **	- 0.0217 **	- 0.0669 ***	- 0.0162 **
	(0.0035)	(0.0085)	(0.0084)	(0.0141)	(0.0052)
Female	0.0040 ***	0.0073 ***	0.0066 **	0.0131 ***	- 0.0014
	(0.0010)	(0.0021)	(0.0024)	(0.0031)	(0.0016)
International	0.0045 ***	0.0043 *	0.0016	- 0.0227 **	0.0006
	(0.0011)	(0.0020)	(0.0033)	(0.0069)	(0.0019)
Thirdparty	- 0.0073 ***	- 0.0130 ***	- 0.0074 ***	- 0.0072 *	- 0.0071 ***
	(0.0009)	(0.0033)	(0.0017)	(0.0036)	(0.0014)
Fees	0.0001	0.0072 ***	- 0.0040 *	- 0.0267 ***	- 0.0029 **
	(0.0008)	(0.0019)	(0.0018)	(0.0035)	(0.0011)
N	1285	182	305	143	655

Tab. 6: Truncated regression results

Notes: Results were obtained from 1 000 bootstrap repetitions. Constants as well as time and country dummies are included but not reported. Bootstrap standard errors are in parentheses. * p < 0.05, ** p < 0.01, and *** p < 0.001.

that neither a negative nor significant relation between the share of female full professors and efficiency can be confirmed.

So far, we have confined attention to effects that, although significant according to the pooled model, do not withstand cluster-specific examination. In addition to this dimension of effect heterogeneity, there is a second group of variables whose influence on efficiency might be overlooked without further scrutiny. Among them, the share of student fees certainly stands out. While the pooled model rejects any notable impact, we observe significant coefficients of $Fees$ in each cluster. More importantly, universities that rely more heavily on student fees are considered more efficient in the physical, health, and general cluster. On the contrary, the opposing relation is found in the social cluster. A similar pattern, though to a lesser extent, becomes visible with regard to regional gross domestic product per capita (GDP) and the indicator for multisite institutions ($Multisite$); that is, coefficients turn significant only in a single cluster.

To draw an interim conclusion, several efficiency drivers appear to differ in relevance between subject clusters. Yet some variables show consistent effects. More specifically, our analysis indicates that efficiency is in general positively related to the share of full professors ($Prof$), the share of third-party funding ($Thirdparty$), and university size ($Size$). A closer look at the magnitude of these coefficient estimates further reveals their economic significance. On average, we would expect inefficiency to decrease in a range between 0.7 and 1.3 percent if the share of third-party funding increased by 1 percentage point. In comparison, raising the share of full professors by an equal margin should lower inefficiency by 1.6 to 6.7 percent. It is worth noting, however, that the latter adjustment might require higher efforts given that personnel structures are supposedly less flexible than revenue compositions.[12] Turning to the impact of institutional size on inefficiency, we estimate point elasticities between -0.10 and -0.40. Despite some variation in effect sizes, we hereby provide evidence for economies of scale in higher education and additionally infer that avenues for efficiency improvement exist on both the personnel and financial level.

7 Robustness Analysis

Within this chapter, we provide further evidence probing the robustness of our results. Specifically, we report a series of model checks that involve variations in peer-group construction, output selection, and regression design.

The initial clustering solution marks the starting point for these analyses. As a first step, we assess the quality of this solution by determining silhouette coefficients for each university. Following Rousseeuw (1987), silhouette coefficients indicate how well (data) objects have been classified by a given partitioning. In more concrete terms, they are derived by comparing an object's proximity to its cluster members with the proximity to the members of its neighbouring cluster, i.e. the cluster with the highest proximity among those the object is not part of. In general, silhouette coefficients can range between -1 and 1 with higher values denoting stronger structures. Consistent with the K-means algorithm, we rely on squared Euclidean distance in subject space to measure proximity between universities. Silhouette coefficients are illustrated in descriptive form in Table 7 and depicted graphically in Appendix D. Two observations stand out from these displays. Firstly, each cluster is characterised by an average silhouette coefficient

[12] Within our regression sample, $Prof$ indeed shows less variation than $Thirdparty$, which is reflected by standard deviations of 2.9 and 13.3 percentage points, respectively.

higher than 0.50, which is commonly referred to as a threshold for reasonable cluster structures. Secondly, however, some universities with silhouette coefficients close to 0 appear to be classified rather vaguely.

Cluster	*N*	Min	P25	P50	Mean	P75	Max
Social	57	-0.123	0.341	0.666	0.535	0.758	0.819
Physical	140	0.105	0.557	0.819	0.696	0.859	0.882
Health	49	-0.042	0.289	0.661	0.524	0.744	0.797
General	204	0.024	0.456	0.611	0.575	0.725	0.828
Sample	450	-0.123	0.470	0.665	0.602	0.783	0.882

Tab. 7: Summary statistics on silhouette coefficients by cluster

The second observation hardly comes as a surprise. While we find the European university landscape to feature four subject clusters, it is to be expected that not all universities fit into this classification. Some institutions obviously occupy niches, which suggests that cluster boundaries are partly fluid. As a consequence, comparing universities to their clusters members could be suboptimal (in some cases). Instead, certain universities, especially those near the boundaries, might possess relevant peers outside their own cluster. We thus extend our approach by constructing tailored peer-groups for each university, which are not bound to cluster affiliation but solely based on proximity in subject space. The advantage of this approach, which we term nearest neighbourhood approach, clearly lies in greater homogeneity. It does, however, imply performing 450 bootstrap DEA estimations per year and therefore requires significantly more computing resources.

Cluster	Social Peers	Physical Peers	Health Peers	General Peers	Δ Distance
Social	67.95%	1.19%	3.41%	27.44%	-32.41%
Physical	0.24%	83.00%	0.00%	16.76%	-20.87%
Health	3.27%	0.00%	61.05%	35.67%	-34.92%
General	2.68%	11.60%	4.47%	81.24%	-25.52%

Tab. 8: Average peer-group composition by cluster, nearest neighbourhood approach
Notes: Δ Distance refers to the change in average squared Euclidean distance between peers resulting from peer-group construction free of cluster constraints.

Peer-group compositions based on our modified approach are reported in Table 9. For reasons of comparability, we hereby stick to identical peer-group sizes as in our baseline model, so that universities in the social cluster, for instance, are assessed relative to their 56 closest peers. The average distance between peers is reduced by a substantial margin of 21 to 35% as we switch to the nearest neighbourhood approach mainly because general cluster universities frequently enhance peer-groups of universities from specialised clusters. Bias-corrected efficiency scores are then estimated based on these individual peer-groups and regressed on our set of covariates giving rise to the results in Table 9. In line with chapter 6, we again observe considerable effect variation across subject fields and find efficiency drivers to remain (in)significant in the majority of cases. Yet some previous findings need to be refined. More specifically, institutional size and the share of third-party funds are no longer generally linked to higher efficiency as these inferences cannot be drawn in the health cluster. Similarly, we observe the share of full professors to impact efficiency less clearly within the social cluster (p-value of 0.07).

Variable	**Social Cluster**	**Physical Cluster**	**Health Cluster**	**General Cluster**
Natural logarithm of bias-corrected efficiency score as dependent variable				
ln(GDP)	- 0.0819	≙ 0.1000	≙ 0.0935	≙ - 0.0687 *
	(0.0587)	(0.0746)	(0.0679)	(0.0350)
Multisite	0.1443 *	≙ 0.1031 *	0.2350 *	≙ 0.0087
	(0.0653)	(0.0506)	(0.1012)	(0.0208)
Hospital	≙ - 0.3071	≙ - 0.0927	≙ 0.3690 ***	≙ 0.0887 ***
	(0.1605)	(0.0625)	(0.0578)	(0.0216)
ln(Size)	≙ - 0.2254 ***	≙ - 0.3967 ***	- 0.0316	≙ - 0.2352 ***
	(0.0597)	(0.0262)	(0.0425)	(0.0187)
Herfindahl	≙ 0.0003	≙ - 0.0033 **	≙ - 0.0005	≙ - 0.0027
	(0.0027)	(0.0011)	(0.0015)	(0.0028)
Prof	- 0.0199	≙ - 0.0295 ***	≙ - 0.0398 **	≙ - 0.0175 ***
	(0.0112)	(0.0086)	(0.0124)	(0.0051)
Female	- 0.0007	≙ 0.0094 ***	≙ 0.0135 ***	≙ - 0.0018
	(0.0029)	(0.0024)	(0.0029)	(0.0016)
International	0.0006	≙ 0.0029	≙ - 0.0239 ***	≙ 0.0008
	(0.0027)	(0.0034)	(0.0063)	(0.0018)
Thirdparty	≙ - 0.0125 **	≙ - 0.0082 ***	- 0.0040	≙ - 0.0067 ***
	(0.0039)	(0.0017)	(0.0030)	(0.0014)
Fees	≙ 0.0111 ***	≙ - 0.0043 *	≙ - 0.0208 ***	≙ - 0.0022 *
	(0.0025)	(0.0019)	(0.0028)	(0.0011)
N	182	305	143	655

Tab. 9: Truncated regression results, nearest neighbourhood approach
Notes: Results were obtained from 1 000 bootstrap repetitions. Constants as well as time and country dummies are included but not reported. ≙ marks coefficient estimates that stay either significant or insignificant compared to Table 6. Bootstrap standard errors are in parentheses. * p < 0.05, ** p < 0.01, and *** p < 0.001.

An additional robustness check refers to altering the concept of university efficiency. To account for increasing levels of technology transfer activities, we now opt for granted patents (instead of publications) to complement citations and graduates as a third output component.[13] As pointed out by Geuna and Nesta (2006), patenting efforts tend to be concentrated in the areas of life sciences and technology, which could partially explain why many European universities did not obtain any patents in the past (OECD, 2003). Our data generally confirm this picture as we find patent engagement to vary considerably across clusters and, besides, note that a number of universities received zero patents. To ensure a comparison of appropriate peers, we thus confine attention to a subgroup of our sample that received at least one patent in every year between 2011 and 2014. Overall, this leaves us with 281 universities with measurable pursuit of applied forms of research output.

All of these universities serve as potential peers as part of the nearest neighbourhood approach. The subsequent regression, however, requires reasonable sample sizes, which leads us to focus on the physical cluster with 79 and the general cluster with 163 universities.[14] Results are documented in columns 4 and 5 of Table 10. With regard to the consistent effects reported earlier, we again detect a significant positive relation between efficiency and both university size and third-party funding shares. In contrast, the share of full professors becomes insignificant. Interestingly, there a different reasons for this finding. In fact, it appears to be attributable to model shift in the physical cluster but to sample composition in the general cluster (see columns 2 and 3). In other words, a higher share of full professors neither improves efficiency in the publication nor in the patent model if we review patent-active general cluster universities. One might assume that these universities rely more on well-run administrations as they tend to be larger and supposedly more complex. In comparison, patent-active universities in the physical cluster seem to benefit from higher shares of full professors as long as we refer to an efficiency concept that builds upon publications instead of patents.

From the standpoint of generalisability, these additional checks allow us to conclude that institutional size and the ability to seek external funding are the main factors to impact university efficiency. With the exception of the health domain, both variables are consistently identified as efficiency-enhancing. To allay concerns about the direction of causality, and to account for possibly delayed effects, further regression analyses with time-lagged covariates are presented in Appendix E. Irrespective of model choice, we find the stated interpretation encouraged by these estimations.

[13] Patent records were derived from Scopus, which contains data from five major patent offices (Elsevier, 2017).

[14] Cluster sizes still constitute the reference points, so that groups of 79 and 163 are used to assess universities from the physical and general cluster, respectively.

Variable	**Physical Cluster** *Publication Model*	**General Cluster** *Publication Model*	**Physical Cluster** *Patent Model*	**General Cluster** *Patent Model*
Natural logarithm of bias-corrected efficiency score as dependent variable				
ln(GDP)	≙ - 0.0376	≙ - 0.1691 ***	0.1564	- 0.2059 ***
	(0.0810)	(0.0404)	(0.1080)	(0.0458)
Multisite	≙ 0.1681 **	≙ - 0.0454	0.1251	- 0.0960 ***
	(0.0608)	(0.0257)	(0.0795)	(0.0286)
Hospital	≙ 0.0077	≙ 0.1582 ***	0.1342	0.1870 ***
	(0.0553)	(0.0261)	(0.0758)	(0.0286)
ln(Size)	≙ - 0.4847 ***	≙ - 0.2863 ***	- 0.5433 ***	- 0.2946 ***
	(0.0417)	(0.0255)	(0.0555)	(0.0280)
Herfindahl	≙ - 0.0048 **	≙ - 0.0018	0.0015	0.0010
	(0.0017)	(0.0035)	(0.0019)	(0.0038)
Prof	≙ - 0.0380 **	- 0.0005	- 0.0230	- 0.0075
	(0.0129)	(0.0061)	(0.0175)	(0.0066)
Female	≙ 0.0158 ***	≙ 0.0002	0.0127 *	0.0003
	(0.0043)	(0.0021)	(0.0058)	(0.0022)
International	≙ 0.0020	0.0052 *	0.0077	0.0052 *
	(0.0036)	(0.0021)	(0.0047)	(0.0023)
Thirdparty	≙ - 0.0132 ***	≙ - 0.0072 ***	- 0.0167 ***	- 0.0083 ***
	(0.0026)	(0.0016)	(0.0034)	(0.0018)
Fees	- 0.0058	- 0.0011	- 0.0070	- 0.0012
	(0.0037)	(0.0013)	(0.0048)	(0.0015)
N	176	534	176	534

Tab. 10: Truncated regression results, nearest neighbourhood approach

Notes: Results were obtained from 1 000 bootstrap repetitions. Constants as well as time and country dummies are included but not reported. ≙ marks coefficient estimates within the publication model that stay either significant or insignificant compared to Table 6. Both models are estimated on identical samples, i.e. the group of patent-active universities in each cluster. Bootstrap standard errors are in parentheses. * p < 0.05, ** p < 0.01, and *** p < 0.001.

8 Conclusion

The present paper studies the relative efficiency of 450 European universities between 2011 and 2014. Our approach is built on the notion that the higher education landscape in Europe is too diverse to be considered one homogeneous peer-group. In particular, differences in subject focus prove indicative for a wide range of institutional characteristics. We uncover these systematic patterns by means of clustering techniques and hereby identify four groups of universities that either possess a balanced subject profile or lay clear emphasis on social sciences, physical sciences, or health sciences. Given that efficiency estimation naturally relies on relative assessment, it is important to differentiate between these distinct groups. Otherwise, one would run the risk of comparing apples and pears. To illustrate this point, we find health cluster universities to incur expenditure per student levels that are, on average, almost four times higher than for social cluster universities.

We address homogeneity concerns firstly by employing intra-cluster efficiency frontiers. In an extension to this approach, we secondly construct individual peer-groups for each university based on subject space proximity. With bias-corrected efficiency scores at hand, we direct attention to potential efficiency drivers, which are investigated within a subsequent regression analysis. It becomes evident that different, partly opposing, mechanisms are in play depending on the cluster under review. Yet institutional size and the ability to seek external funding are largely found to be efficiency-enhancing. Apart from the health cluster, inefficiency is expected to fall by 6.7 to 16.7% if third-party funding shares increased by 10 percentage points and by 1.7 to 5.4% if universities were to expand their capacities by 10%.

Overall, this paper underlines the high degree of diversity in Europe’s higher education sector and provides a framework for further in-depth studies. However, our analyses are not without limitations. Incorporating teaching quality would certainly complement our efficiency perception, yet it is hard to think of reliable measures for this domain. Despite the time-lag regression design, it would also be beneficial to adopt additional methods dedicated to causal inference. Lastly, future research may emphasise the distinction between private and public sources of external funding in order to broaden the understanding of university efficiency beyond the findings presented in this study.

References

Abbott, M., & Doucouliagos, C. (2003). The efficiency of Australian universities: A data envelopment analysis. *Economics of Education Review*, *22*(1), 89–97.

Adams, J. (2005). Early citation counts correlate with accumulated impact. *Scientometrics*, *63*(3), 567–581.

Agasisti, T., & Haelermans, C. (2016). Comparing Efficiency of Public Universities among European Countries: Different Incentives Lead to Different Performances. *Higher Education Quarterly*, *70*(1), 81–104.

Agasisti, T., & Johnes, G. (2009). Beyond frontiers: Comparing the efficiency of higher education decision-making units across more than one country. *Education Economics*, *17*(1), 59–79.

Agasisti, T., & Pérez-Esparrells, C. (2010). Comparing efficiency in a cross-country perspective: The case of Italian and Spanish state universities. *Higher Education*, *59*(1), 85–103.

Agasisti, T., & Pohl, C. (2012). Comparing German and Italian public universities: Convergence or divergence in the higher education landscape? *Managerial and Decision Economics*, *33*(2), 71–85.

Agasisti, T., & Salerno, C. (2007). Assessing the cost efficiency of Italian universities. *Education Economics*, *15*(4), 455–471.

Agasisti, T., & Wolszczak-Derlacz, J. (2016). Exploring efficiency differentials between Italian and Polish universities, 2001-11. *Science and Public Policy*, *43*(1), 128–142.

Ahn, T., Charnes, A., & Cooper, W. W. (1988). Some statistical and DEA evaluations of relative efficiencies of public and private institutions of higher learning. *Socio-Economic Planning Sciences*, *22*(6), 259–269.

Ahn, T., & Seiford, L. M. (1993). Sensitivity of DEA to the Models and Variable Sets in a Hyphothesis Test Setting: The Efficiency of University Operations. *Creative and Innovative Approaches to the Sciences of Management*, 191–210.

Aizenman, J., & Kletzer, K. (2011). The life cycle of scholars and papers in economics - the “citation death tax.” *Applied Economics*, *43*(27), 4135–4148.

Athanassopoulos, A. D., & Shale, E. (1997). Assessing the comparative efficiency of higher education institutions in the UK by the means of data envelopment analysis. *Education Economics*, *5*(2), 117–134.

Beasley, J. E. (1990). Comparing university departments. *Omega*, *18*(2), 171–183.

Bolli, T., Olivares, M., Bonaccorsi, A., Daraio, C., Aracil, A. G., & Lepori, B. (2016). The differential effects of competitive funding on the production frontier and the efficiency of universities. *Economics of Education Review*, *52*, 91–104.

Bonaccorsi, A., & Daraio, C. (2009). Characterizing the European university system: a preliminary classification using census microdata. *Science and Public Policy*, *36*(10), 763–775.

Bonaccorsi, A., Daraio, C., & Simar, L. (2006). Advanced indicators of productivity of universities. An application of robust nonparametric methods to Italian data. *Scientometrics*, *66*(2), 389–410.

Bonaccorsi, A., Secondi, L., Setteducati, E., & Ancaiani, A. (2014). Participation and commitment in third-party research funding: Evidence from Italian Universities. *Journal of Technology Transfer*, *39*(2), 169–198.

Breu, T. M., & Raab, R. L. (1994). Efficiency and perceived quality of the nation's "top 25" National Universities and National Liberal Arts Colleges: An application of data envelopment analysis to higher education. *Socio-Economic Planning Sciences*, *28*(1), 33–45.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods*, *3*(1), 1–27.

Coelli, T. J. (1996). *Assessing the performance of Australian universities using data envelopment analysis*. University of New England, Department of Econometrics.

Cohn, E., Rhine, S. L. W., & Santos, M. C. (1989). Institutions of higher education as multi-product firms: Economies of scale and scope. *The Review of Economics and Statistics*, 284–290.

Daraio, C., Bonaccorsi, A., & Simar, L. (2015a). Efficiency and economies of scale and specialization in European universities: A directional distance approach. *Journal of Informetrics*, *9*(3), 430–448.

Daraio, C., Bonaccorsi, A., & Simar, L. (2015b). Rankings and university performance: A conditional multidimensional approach. *European Journal of Operational Research*, *244*(3), 918–930.

De Witte, K., & López-Torres, L. (2017). Efficiency in education: a review of literature and a way forward. *Journal of the Operational Research Society*, *68*(4), 339–363.

Dundar, H., & Lewis, D. R. (1995). Departmental productivity in American universities: Economies of scale and scope. *Economics of Education Review*, *14*(2), 119–144.

Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., & Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research*, *132*(2), 245–259.

Filippini, M., & Lepori, B. (2007). Cost structure, economies of capacity utilization and scope in Swiss higher education institutions. *Universities and Strategic Knowledge Creation: Specialization and Performance in Europe*, 272–304.

Geuna, A., & Nesta, L. J. J. (2006). University patenting and its effects on academic research: The emerging European evidence. *Research Policy*, *35*(6), 790–807.

Gulbrandsen, M., & Smeby, J. C. (2005). Industry funding and university professors' research performance. *Research Policy*, *34*(6), 932–950.

Hanke, M., & Leopoldseder, T. (1998). Comparing the efficiency of austrian universitiesa data envelopment analysis application. *Tertiary Education and Management*, *4*(3), 191–197.

Hornbostel, S. (2001). Third party funding of German universities. An indicator of research activity? *Scientometrics*, *50*(3), 523–537.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Johnes, J. (1990). Unit costs: Some explanations of the differences between UK universities. *Applied Economics*, *22*(7), 853–862.

Johnes, J. (2006). Data envelopment analysis and its application to the measurement of efficiency in higher education. *Economics of Education Review*, *25*(3), 273–288.

Johnes, J., & Johnes, G. (1995). Research funding and performance in U.K. University Departments of Economics: A frontier analysis. *Economics of Education Review*, *14*(3), 301–314.

Johnes, J., & Yu, L. (2008). Measuring the research performance of Chinese higher education institutions using data envelopment analysis. *China Economic Review*, *19*(4), 679–696.

Joumady, O., & Ris, C. (2005). Performance in European higher education: A non-parametric production frontier approach. *Education Economics*, *13*(2), 189–205.

Katharaki, M., & Katharakis, G. (2010). A comparative assessment of Greek universities' efficiency using quantitative analysis. *International Journal of Educational Research*, *49*(4–5), 115–128.

Kempkes, G., & Pohl, C. (2010). The efficiency of German universities - some evidence from nonparametric and parametric methods. *Applied Economics*, *42*(16), 2063–2079.

Leitner, K. H., Prikoszovits, J., Schaffhauser-Linzatti, M., Stowasser, R., & Wagner, K. (2007). The impact of size and specialisation on universities' department performance: A DEA analysis applied to Austrian universities. *Higher Education*, *53*(4), 517–538.

Lepori, B., Baschung, L., & Probst, C. (2010). Patterns of subject mix in higher education institutions: A first empirical analysis using the aquameth database. *Minerva*, *48*(1), 73–99.

Lepori, B., Bonaccorsi, A., Daraio, A., Daraio, C., Gunnes, H., Hovdhaugen, E., … Wagner-Schuster, D. (2016). *Establishing a European Tertiary Education Register*. Publications Office of the European Union.

Makles, A. (2012). Stata tip 110: How to get the optimal k-means cluster solution. *Stata Journal*, *12*(2), 347–351.

Martin, B. R., & Irvine, J. (1983). Assessing basic research. Some partial indicators of scientific progress in radio astronomy. *Research Policy*, *12*(2), 61–90.

McMillan, M. L., & Datta, D. (1998). The relative efficiencies of Canadian universities: a DEA perspective. *Canadian Public Policy/Analyse de Politiques*, 485–511.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179.

Moed, H. F., Burger, W. J. M., Frankfort, J. G., & Van Raan, A. F. J. (1985). The use of bibliometric data for the measurement of university research performance. *Research Policy*, *14*(3), 131–149.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228.

Necmi K. Avkiran. (2001). Investigating technical and scale efficiencies of Australian Universities through data envelopment analysis. *Socio-Economic Planning Sciences*, *35*, 57–80.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, *66*(1), 81–100.

OECD. (2003). *Turning science into business: Patenting and licensing at public research organisations*. OECD Publishing.

Piro, F. N., Aksnes, D. W., & Rørstad, K. (2013). A macro analysis of productivity differences across fields: Challenges in the measurement of scientific publishing. *Journal of the American Society for Information Science and Technology*, *64*(2), 307–320.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward. *Proceedings of the National Academy of Sciences of the United States of America*, 1–5.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65.

Sagarra, M., Mar-Molinero, C., & Agasisti, T. (2017). Exploring the efficiency of Mexican universities: Integrating Data Envelopment Analysis and Multidimensional Scaling. *Omega (United Kingdom)*, *67*, 123–133.

Sarrico, C. S., Hogan, S. M., Dyson, R. G., & Athanassopoulos, A. D. (1997). Data envelopment analysis and university selection. *Journal of the Operational Research Society*, *48*(12), 1163–1177.

Shin, J. C. (2009). Classifying higher education institutions in Korea: A performance-based approach. *Higher Education*, *57*(2), 247–266.

Shin, J. C., & Cummings, W. K. (2010). Multilevel analysis of academic publishing across disciplines: Research preference, collaboration, and time on research. *Scientometrics*, *85*(2), 581–594.

Sickles, R. C., & Zelenyuk, V. (2019). *Measurement of Productivity and Efficiency*. Cambridge University Press.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. *Monographs on Statistics and Applied Probability, London: Chapman and Hall, 1986*.

Simar, L., & Wilson, P. W. (1998). Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models. *Management Science*, *44*(1), 49–61.

Simar, L., & Wilson, P. W. (2000). A general methodology for bootstrapping in non-parametric frontier models. *Journal of Applied Statistics*, *27*(6), 779–802.

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.

Smith, J. P., & Naylor, R. A. (2001a). Determinants of degree performance in UK Universities: a statistical analyses. *Oxford Bulletin of Economics and Statistics*, *63*(1), 29–60.

Smith, J. P., & Naylor, R. A. (2001b). Dropping out of university: A statistical analysis of the probability of withdrawal for UK university students. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *164*(2), 389–405.

Stanley, G., & Reynolds, P. (1994). Similarity grouping of Australian universities. *Higher Education*, *27*(3), 359–366.

Taylor, B., & Harris, G. (2004). Relative efficiency among South African universities: A data envelopment analysis. *Higher Education*, *47*(1), 73–89.

Teixeira, P. N., Rocha, V., Biscaia, R., & Cardoso, M. F. (2012). Competition and diversity in higher education: An empirical approach to specialization patterns of Portuguese institutions. *Higher Education*, *63*(3), 337–352.

Thanassoulis, E., Kortelainen, M., Johnes, G., & Johnes, J. (2011). Costs and efficiency of higher education institutions in England: A DEA analysis. *Journal of the Operational Research Society*, *62*(7), 1282–1297.

Tierney, M. L. (1980). An estimate of departmental cost functions. *Higher Education*, *9*(4), 453–468.

Tomkins, C., & Green, R. (1988). An experiment in the use of data envelopment analysis for evaluating the efficiency of UK university departments of accounting. *Financial Accountability & Management*, *4*(2), 147–164.

Valadkhani, A., & Worthington, A. (2006). Ranking and clustering Australian university research performance, 1998--2002. *Journal of Higher Education Policy and Management*, *28*(2), 189–210.

Walker, I., & Zhu, Y. (2013). *The impact of university degrees on the lifecycle of earnings: some further analysis*.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: An empirical analysis. *Scientometrics*, *87*(3), 467–481.

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, *94*(3), 851–872.

Wolszczak-Derlacz, J. (2017). An evaluation and explanation of (in)efficiency in higher education institutions in Europe and the U.S. with the application of two-stage semi-parametric DEA. *Research Policy*, *46*(9), 1595–1605.

Wolszczak-Derlacz, J., & Parteka, A. (2011). Efficiency of European public higher education institutions: a two-stage multicountry approach. *Scientometrics*, *89*(3), 887.

Worthington, A. C. (2002). An Empirical Survey of Frontier Efficiency Measurement Techniques in Education. *Education Economics*, *9*(3), 245–268.

Zimmerman, S. D., & Altonji, J. G. (2018). The Costs of and Net Returns to College Major. *National Bureau of Economic Research*.

Zoghbi, A. C., Rocha, F., & Mattos, E. (2013). Education production efficiency: Evidence from Brazilian universities. *Economic Modelling*, *31*(1), 94–103.

Appendix A. Evaluating the Number of Clusters

In order to test our decision for four clusters, we consider a second heuristic, i.e. the Calisnki-Harabasz index (also termed pseudo-F). According to a comparative study by Milligan and Cooper (1985), this index performed best among 30 stopping rules and has since become a standard procedure in clustering analysis. Technically, it combines compactness and separation in its formula, with the latter aspect being an addition to our baseline approach. (Separation reflects how distinct clusters are from each other whereas compactness relates to the similarity within clusters).

Clusters	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Pseudo-F	319.6	351.8	380.1	366.1	356.7	356.3	359.5	362.1	366.3	364.9	363.0
Best Case	0	0	874	2	2	12	0	0	100	6	4

Tab. A.1: Evaluation based on Calinski-Harabasz approach

Notes: Pseudo-F values are averaged over 1 000 replications of K-means with random starting centres. "Best Case" indicates how often a clustering solution was recommended based on this criterion.

Consistent with our methodology outlined in chapter 3.2, we ran the K-means algorithm 1 000 times based on the same sequence of random starting centres. The results reported in Table A.1 again favour four clusters given that the average Calinski-Harabasz index reaches a maximum for this configuration. In fact, we observe a four-cluster solution being recommended in 87.4% of our replications. On these grounds, our initial choice can be confirmed.

Appendix B. Citation Window Analysis

Within our model specification, we employ citations as an indicator of research impact. Since our data covers a citation window of three years, it is worth discussing if this time span complies with long-term impact. For this purpose, we traced our universities back to 1996 and compiled a new dataset consisting of 184 766 publications.

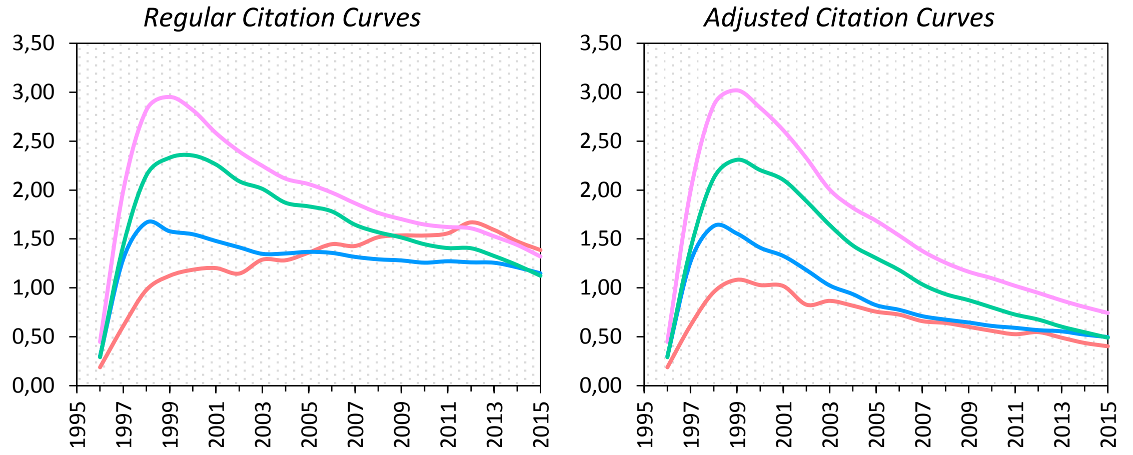

Fig. B.1: Regular and adjusted citation curves for publications from 1996 by field
Notes: Colouring refers to life (rose), health (green), physical (blue), and social (red) sciences.

As a first step, we calculated citation curves, i.e. average annual citation counts, for the subset of publications with at least one citation over the 20-year period from 1996 to 2015. This subset comprises 86.18% of our initial data. As can be seen in the left-hand panel of Figure B.1, the curves of life, health, and physical sciences follow a similar shape, peaking between 1998 and 2000, followed by a steady decline. Citation counts in social sciences on the other hand continue to rise until 2012. From a theoretical standpoint, one might argue that this difference is primarily due to a slower pace of theoretical development (Nederhof, 2006). Interestingly, however, we discover this pattern to be largely attributable to a higher growth rate of social sciences within the Scopus database. Once we deflate citation counts by field-specific growth rates, social sciences clearly becomes less of an exemption as illustrated by the right-hand panel of Figure B.1.[15]

Our graphical depiction further reveals citations not only to differ in absolute numbers but also regarding the way they mature. This finding could potentially raise concerns about the accuracy of using short-term citations as a predictor for long-term citations. In order to test this relation, we examine correlations between cumulative citation counts over

[15] Following Aizenman and Kletzer (2011), citations counts are divided by a time-varying index, defined as the number of publications in a given year relative to the number of publications in our base year (1996). Of course, indices are calculated separately for each field. Moreover, it should be noted that our adjustment is not based on the full Scopus database but on a comprehensive subset of 15.6 million publications.

increasing time spans each starting in 1996 and total citation counts. Wang (2013) rightly points out that citation counts are far from being normally distributed, so that Spearman correlations are expected to be most reliable. Yet we also report Pearson correlations to allow comparison with previous studies, for instance by Adams (2005) or Waltman et al. (2011). Results are presented in Table B.1.

Year	**Life Sciences**		**Social Sciences**		**Physical Sciences**		**Health Sciences**	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
1996	0.345	0.350	0.313	0.280	0.141	0.321	0.354	0.333
1997	0.550	0.657	0.510	0.533	0.261	0.616	0.561	0.643
1998	0.647	0.790	0.652	0.693	0.341	0.746	0.655	0.789
1999	0.703	0.854	0.722	0.781	0.403	0.813	0.715	0.859
2000	0.745	0.891	0.772	0.835	0.465	0.857	0.763	0.898
2001	0.781	0.915	0.813	0.869	0.532	0.887	0.805	0.923
2002	0.813	0.932	0.840	0.893	0.601	0.909	0.840	0.940
2003	0.841	0.946	0.870	0.914	0.667	0.927	0.869	0.953
2004	0.864	0.957	0.895	0.931	0.726	0.941	0.894	0.963
2005	0.887	0.966	0.916	0.945	0.782	0.953	0.917	0.971
2006	0.907	0.973	0.936	0.956	0.832	0.963	0.936	0.978
2007	0.924	0.980	0.951	0.965	0.875	0.971	0.950	0.983
2008	0.940	0.985	0.964	0.973	0.909	0.978	0.963	0.987
2009	0.956	0.989	0.975	0.980	0.939	0.983	0.974	0.991
2010	0.969	0.992	0.983	0.985	0.960	0.988	0.983	0.993

Tab. B.1: Correlation between cumulative and total citation counts by field

Notes: Cumulative citation counts span the period from 1996 up to and including the year given in the first column. Total citation counts cover 20 years (1996-2015).

From Table B.1, we can infer that short-term citations vary in their accuracy as a proxy for long-term citations. In social sciences, for instance, it would require 8 years to exceed a Spearman correlation of 0.9 whereas 6 years would be sufficient in life or health sciences. Of course, it is hard to define an acceptable level of correlation. However, we might be in a position to circumvent this question. In fact, our research design is not concerned about correlations on the level of single publications but on an institutional level. By aggregating citations over universities and calculating correlations afterwards for each cluster, we notice a considerable increase in the degree of dependence between initial and overall citations (see Table B.2). Apparently, variation is largely cancelled out when taking an institutional perspective as illustrated by almost perfect correlations. This result then leads us to conclude that relatively short citation windows indeed provide a reliable basis for our study.

Year	**Social Cluster**		**Physical Cluster**		**Health Cluster**		**General Cluster**	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
1996	0.994	0.858	0.974	0.964	0.983	0.972	0.977	0.985
1997	0.994	0.958	0.977	0.979	0.990	0.981	0.983	0.989
1998	0.994	0.975	0.982	0.984	0.991	0.988	0.986	0.991
1999	0.995	0.984	0.984	0.987	0.992	0.990	0.987	0.992
2000	0.995	0.984	0.987	0.988	0.993	0.993	0.989	0.993

Tab. B.2: Correlation between cumulative and total citation counts by cluster

Notes: Cumulative citation counts span the period from 1996 up to and including the year given in the first column. Total citation counts cover 20 years (1996-2015). Citations are aggregated by institutions.

Appendix C. Additional Density Plots

Great Britain, Germany, and Italy constitute the three largest countries in our dataset with 120, 79, and 64 universities, respectively. Given these sample sizes, those countries are most suitable for a cluster-specific comparison based on density estimates. As illustrated in Figure C.1, Great Britain shows high levels of efficiency across all clusters. With the exception of the social cluster, Italy also performs well quite closely resembling Great Britain's distributions. Germany, on the other hand, displays consistently lower efficiency levels. Interestingly, Germany's efficiency estimates are rather evenly distributed thus suggesting that the national landscape appears very heterogeneous from an efficiency standpoint.

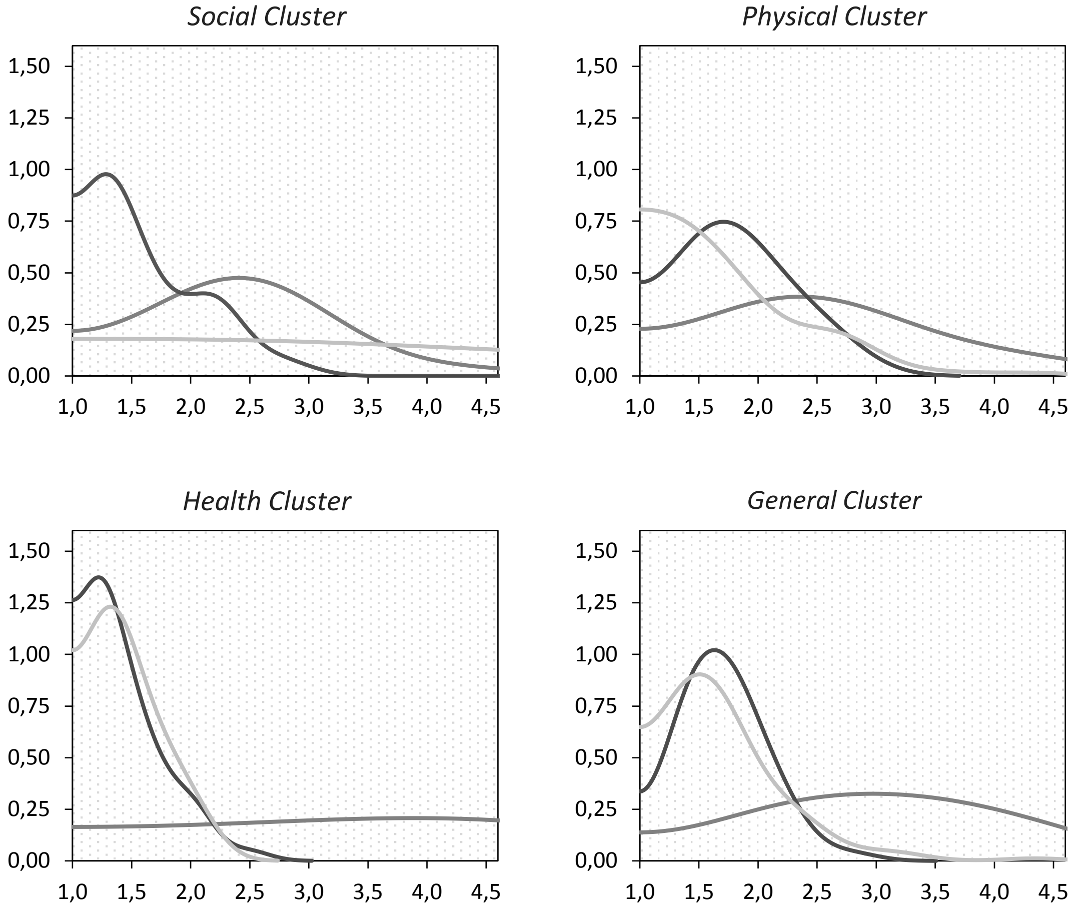

Fig. C.1: Density estimates of bias-corrected efficiency scores by cluster and country
Notes: Great Britain (dark), Germany (medium), and Italy (light) are distinguished by greyscale.

Appendix D. Silhouette Plots

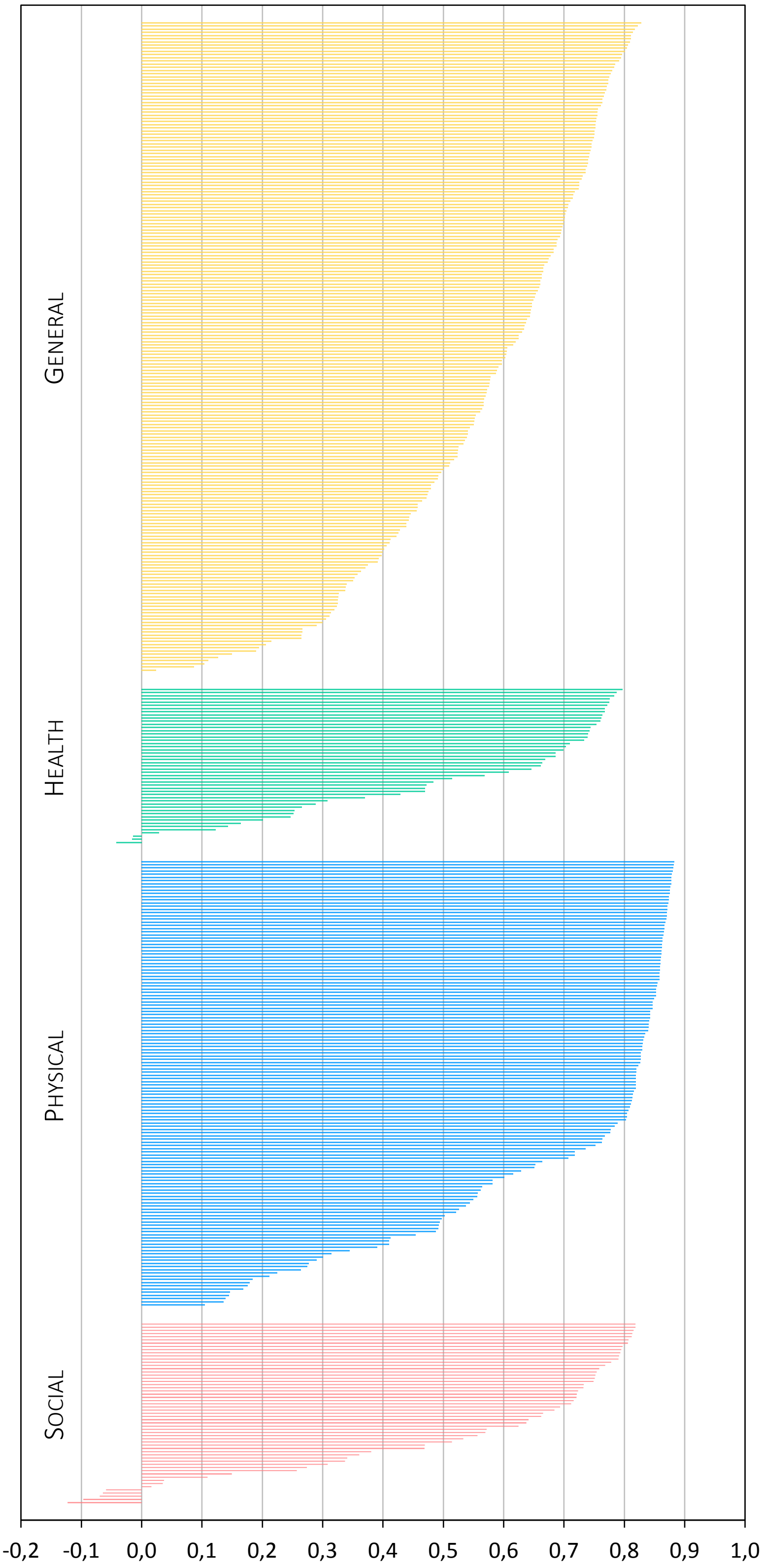

Fig. D.1: Silhouette plots by cluster

Appendix E. Time-Lag Regression Design

Reverse causality is well known to hinder clear inference. In the context of this study, it may arise if universities become more successful in competing for third-party funds as a result of increased efficiency. We attempt to avoid this problem by using time-lagged variables. The idea behind this approach is that, within a given year, funding structures could be affected by efficiency; however, it is unlikely for past funding structures to be subject to the same problem. As this reasoning can also be applied to other variables, we universally employ a one-year time lag. Results of both the clustering and nearest neighbourhood approach are reported below. Overall, it appears that our main results are largely confirmed.

Variable	**Social Cluster**	**Physical Cluster**	**Health Cluster**	**General Cluster**
Natural logarithm of bias-corrected efficiency score as dependent variable				
ln(GDP)	- 0.1559 **	0.1396 *	0.0691	- 0.0679
	(0.0518)	(0.0687)	(0.0897)	(0.0395)
Multisite	0.0492	0.1124 *	0.2030	0.0082
	(0.0594)	(0.0479)	(0.1401)	(0.0234)
Hospital	- 0.2850 *	- 0.0458	0.2840 ***	0.1085 ***
	(0.1333)	(0.0581)	(0.0784)	(0.0255)
ln(Size)	- 0.2105 ***	- 0.3369 ***	- 0.1014	- 0.1621 ***
	(0.0536)	(0.0255)	(0.0599)	(0.0208)
Herfindahl	0.0050 *	- 0.0026 **	- 0.0025	- 0.0041
	(0.0024)	(0.0010)	(0.0019)	(0.0030)
Prof	- 0.0209 *	- 0.0269 **	- 0.0619 ***	- 0.0106
	(0.0086)	(0.0086)	(0.0170)	(0.0056)
Female	0.0074 **	0.0051 *	0.0142 ***	- 0.0041 *
	(0.0024)	(0.0023)	(0.0038)	(0.0016)
International	0.0031	0.0023	- 0.0209 *	0.0001
	(0.0024)	(0.0031)	(0.0087)	(0.0021)
Thirdparty	- 0.0146 ***	- 0.0076 ***	- 0.0088 *	- 0.0075 ***
	(0.0038)	(0.0016)	(0.0041)	(0.0016)
Fees	0.0066 **	- 0.0041 *	- 0.0308 ***	- 0.0030 *
	(0.0024)	(0.0018)	(0.0045)	(0.0013)
N	134	225	106	487

Tab. E.1: Truncated regression results, clustering approach with time lag

Notes: Results were obtained from 1 000 bootstrap repetitions. Constants as well as time and country dummies are included but not reported. The model employs a one-year time lag, i.e. efficiency scores from year t are regressed on explanatory variables from year t-1. Bootstrap standard errors are in parentheses. * p < 0.05, ** p < 0.01, and *** p < 0.001.

Variable	Social Cluster	Physical Cluster	Health Cluster	General Cluster
Natural logarithm of bias-corrected efficiency score as dependent variable				
ln(GDP)	- 0.0661	0.1415 *	- 0.1244	- 0.0652
	(0.0618)	(0.0713)	(0.0798)	(0.0363)
Multisite	0.1631 *	0.1097 *	0.2766 *	0.0076
	(0.0720)	(0.0486)	(0.1286)	(0.0223)
Hospital	- 0.3313	0.0704	0.3884 ***	0.0870 ***
	(0.1708)	(0.0647)	(0.0705)	(0.0249)
ln(Size)	- 0.1798 **	- 0.3396 ***	- 0.0404	- 0.2133 ***
	(0.0638)	(0.0278)	(0.0556)	(0.0208)
Herfindahl	0.0010	- 0.0027 *	0.0003	- 0.0050
	(0.0028)	(0.0011)	(0.0016)	(0.0030)
Prof	- 0.0133	- 0.0349 ***	- 0.0371 *	- 0.0104
	(0.0110)	(0.0088)	(0.0151)	(0.0054)
Female	- 0.0029	0.0071 **	0.0149 ***	- 0.0046 **
	(0.0028)	(0.0025)	(0.0033)	(0.0016)
International	- 0.0012	0.0034	- 0.0240 **	0.0000
	(0.0028)	(0.0033)	(0.0078)	(0.0019)
Thirdparty	- 0.0134 **	- 0.0083 ***	- 0.0044	- 0.0073 ***
	(0.0046)	(0.0018)	(0.0034)	(0.0015)
Fees	0.0106 ***	- 0.0043 *	- 0.0231 ***	- 0.0024
	(0.0028)	(0.0020)	(0.0036)	(0.0013)
N	134	225	106	487

Tab. E.2: Truncated regression results, nearest neighbourhood approach with time lag

Notes: Results were obtained from 1 000 bootstrap repetitions. Constants as well as time and country dummies are included but not reported. The model employs a one-year time lag, i.e. efficiency scores from year t are regressed on explanatory variables from year t-1. Bootstrap standard errors are in parentheses. * p < 0.05, ** p < 0.01, and *** p < 0.001.