QSS

an open access journal

Check for updates

Citation: Abramo, G., & D'Angelo, C. A. (2023). How reliable are unsupervised author disambiguation algorithms in the assessment of research organization performance? *Quantitative Science Studies*, *4*(1), 144–166. https://doi.org/10.1162/qss_a_00236

DOI:
https://doi.org/10.1162/qss_a_00236

Peer Review:
https://www.webofscience.com/api/gateway/wos/peer-review/10.1162/qss_a_00236

Received: 19 July 2022
Accepted: 18 December 2022

Corresponding Author:
Giovanni Abramo
giovanni.abramo@iasi.cnr.it

Handling Editor:
Ludo Waltman

The MIT Press

RESEARCH ARTICLE

How reliable are unsupervised author disambiguation algorithms in the assessment of research organization performance?

Giovanni Abramo[1] and **Ciriaco Andrea D'Angelo**[1,2]

[1]Laboratory for Studies in Research Evaluation, Institute for System Analysis and Computer Science (IASI-CNR), National Research Council of Italy, Rome, Italy

[2]Department of Engineering and Management, University of Rome "Tor Vergata," Rome, Italy

Keywords: author name disambiguation, evaluative scientometrics, FSS, Italy, research assessment, universities

ABSTRACT

Assessing the performance of universities by output to input indicators requires knowledge of the individual researchers working within them. Although in Italy the Ministry of University and Research updates a database of university professors, in all those countries where such databases are not available, measuring research performance is a formidable task. One possibility is to trace the research personnel of institutions indirectly through their publications, using bibliographic repertories together with author names disambiguation algorithms. This work evaluates the goodness-of-fit of the Caron and van Eck, CvE unsupervised algorithm by comparing the research performance of Italian universities resulting from its application for the derivation of the universities' research staff, with that resulting from the supervised algorithm of D'Angelo, Giuffrida, and Abramo (2011), which avails of input data. Results show that the CvE algorithm overestimates the size of the research staff of organizations by 56%. Nonetheless, the performance scores and ranks recorded in the two compared modes show a significant and high correlation. Still, nine out of 69 universities show rank deviations of two quartiles. Measuring the extent of distortions inherent in any evaluation exercises using unsupervised algorithms, can inform policymakers' decisions on building national research staff databases, instead of settling for the unsupervised approaches.

1. INTRODUCTION

The tools of performance assessment play a fundamental role in the strategic planning and analysis of national and regional research systems, member organizations and individuals. At the level of research organizations, assessment serves in identifying fields of strength and weakness, which in turn inform competitive strategies, organizational restructuring, resource allocation, and individual incentive systems. For regions and countries, knowledge of strengths and weaknesses relative to others, and also the comparative performances of one's own research institutions, enables formulation of informed research policies and selective allocation of public funding across fields and institutions. By assessing performance before versus after, institutions and governments can evaluate the impact of their strategic actions and implementation of policy (Karlsson, 2017; Gläser & Laudel, 2016). Performance-based research funding and rewards stimulate improvement of performance. Such assessments also serve in reducing information asymmetries between the suppliers (researchers, institutions, territories)

and the end users of research (companies, students, investors). At the macroeconomic level, this yields twofold beneficial results, resulting in a virtuous circle: a) In selecting research suppliers, users can make more effective choices; and b) suppliers, aiming to attract more users, will be stimulated to improve their research production. The reduction of asymmetric information is also beneficial within the scientific communities themselves, particularly in the face of the increasing challenges of complex interdisciplinary research, by lowering obstacles among prospective partners as they seek to identify others suited for inclusion in team-building.

Over recent years, the stakeholders of research systems have demanded more timely assessment, capable of informing in an ever more precise, reliable, and robust manner (Zacharewicz, Lepori et al., 2019). Bibliometrics, and in particular evaluative bibliometrics, has the great advantage of enabling large-scale research evaluations with levels of accuracy, costs, and timescales far more advantageous than traditional peer-review (Abramo, D'Angelo, & Reale, 2019), as well as possibilities for informing small-scale peer-review evaluations. For years, in view of the needs expressed by policymakers, research managers, and stakeholders in general, scholars have continuously improved the indicators and methods of evaluative bibliometrics. In our opinion, however, the key factor holding bibliometricians back from a great leap forward is the lack of input data, which in almost all nations have been very difficult to assemble.

In all production systems, the comparative performance of any unit is always given by the ratio of outputs to inputs. In the case of research systems, the inputs or production factors consist basically of labor (the researchers) and capital (all resources other than labor, e.g., equipment, facilities, databases, etc.). For any research unit, therefore, comparison to another demands that we are informed of the component researchers, and the resources they draw on for conducting their research. In addition, bias in results would occur unless also informed of the prevailing research discipline of each researcher, because output (new knowledge encoded in publications and the like), all inputs equal, is in part a function of discipline (Sorzano, Vargas et al., 2014; Piro, Aksnes, & Rørstad, 2013; Lillquist & Green, 2010; Sandström & Sandström, 2009; Iglesias & Pecharromán, 2007; Zitt, Ramanana-Rahary, & Bassecoulard, 2005): scholars of blood diseases, for example, publish an average of about five times as much as scholars of legal medicine (D'Angelo & Abramo, 2015). Finally, the measure of the researcher's contribution to each scientific output should also take into account the number of coauthors, and in some cases their position in the author list (Waltman & van Eck, 2015; Abramo, D'Angelo, & Rosati, 2013; Aksnes, Schneider, & Gunnarsson, 2012; Huang, Lin, & Chen, 2011; Gauffriau & Larsen, 2005; van Hooydonk, 1997; Rinia, De Lange, & Moed, 1993).

Yet for many years, regardless of all the above requirements, organizations have regularly published research institution performance rankings that are coauthor, size, and field dependent, among which the most renowned would be the *Academic Ranking of World Universities (ARWU)*,[1] issued by Shanghai Jiao Tong University, and the *Times Higher Education World University Rankings*.[2] Despite the strong distortions in these rankings (Butler, 2010; Dehon, McCathie, & Verardi, 2010; Turner, 2005; van Raan, 2005), many decision-makers persist in giving them serious credit. One of the most recent gestures of the sort came in May 2022, when the British government, intending well for those seeking immigration but without a job offer, offered early-career "High Potential Individuals" the possibility of a visa, subject to graduation within the past five years from an eligible university: meaning any university placing near top of the above—highly distorted—rankings.[3]

[1] https://www.shanghairanking.com (last accessed 15/12/2022).
[2] https://www.timeshighereducation.com/world-university-rankings (last accessed 15/12/2022).
[3] https://www.gov.uk/government/publications/high-potential-individual-visa-global-universities-list.

To get around the obstacle of missing input data, some bibliometricians have seen a solution in the so-called *size independent* indicators of research performance—among these the world-famous mean normalized citation score or MNCS (Waltman, van Eck et al., 2011; Moed, 2010). However, these indicators result in performance scores and ranks that are different from those obtained using other indicators, such as FSS (fractional scientific strength), which do account for inputs, albeit with certain unavoidable assumptions.[4] But the FSS indicator has thus far been applied in only two countries, both with advantages of government records on inputs: extensively, in Italy, for the evaluation of performance at the level of individuals (Abramo & D'Angelo, 2011) and then aggregated at the levels of research field and university (Abramo, D'Angelo, & Di Costa, 2011), and to a lesser extent in Norway, with additional assumptions (Abramo, Aksnes, & D'Angelo, 2020).

For policymakers and administrators, but also all interested others, the question then becomes: "in demanding and/or using large-scale assessments of the positioning of research institution performance, what margin of error is acceptable in the measure of their scores and ranks?" To give an idea of the potential margins of error, a comparison of research-performance scores and ranks of Italian universities by MNCS and FSS revealed that 48.4% of universities shifted quartiles under these two indicators, and that 31.3% of universities in the top quartile by FSS fell into lower quartiles by MNCS (Abramo & D'Angelo, 2016c).

Italy is an almost completely unique case in the provision of the data on the research staff at universities, as necessary for institutional performance evaluation. Here, at the close of each year, the Ministry of University and Research (MUR) updates a database of all university faculty members, listing the first and last names of each researcher, their gender, institutional affiliation, field classification and academic rank.[5] The Norwegian Research Personnel Register also offers a useful database of statistics,[6] including notation of the capital cost of research per person-year aggregated at area level, based on regular reports from the institutions to the Nordic Institute for Studies in Innovation, Research and Education (NIFU).[7]

The challenge that practitioners are facing is then how to apply output-input indicators of research performance aligned with microeconomic theory of production (like FSS), in all those countries where databases of personnel are not maintained. One possibility is to trace the research personnel of the institutions indirectly, through their publications, using bibliographic repertories such as Scopus or Web of Science (WoS), and referring exclusively to bibliometric metadata, apply algorithms for disambiguation of authors' names and reconciling of the institutions' names.

Computer scientists and bibliometricians have developed several unsupervised algorithms for disambiguation, at national and international levels (Rose & Kitchin, 2019; Backes, 2018a; Hussain & Asghar, 2018; Zhu, Wu et al., 2017; Liu, Doğan et al., 2014; Caron & van Eck, 2014; Schulz, Mazloumian et al., 2014; Wu, Li et al., 2014; Wu & Ding, 2013; Cota, Gonçalves, & Laender, 2007). The term *unsupervised* signifies that the algorithms operate without manually labelled data, instead approaching the author-name disambiguation problem as a clustering task, where each cluster would contain all the publications written by a specific author. Tekles and Bornmann (2020), using a large validation set containing more than one million author mentions, each annotated with a Researcher ID (an identifier maintained by the researchers), compared a set of such unsupervised disambiguation approaches. The best performing

[4] A thorough explanation of the theory and assumptions underlying FSS can be found in Abramo et al. (2020).
[5] https://cercauniversita.cineca.it/php5/docenti/cerca.php, last accessed 15/12/2022.
[6] https://www.nifu.no/en/statistics-indicators/4897-2/, last accessed 15/12/2022.
[7] https://www.foustatistikkbanken.no/nifu/?language=en, last accessed 15/12/2022.

algorithm resulted as the one by Caron and van Eck (2014), hereinafter "CvE." As discussed above, however, the conduct of performance comparisons at organizational level requires more than just precision in unambiguously attributing publications to each author. At that point we also need precise identification of the research staff of each organization,[8] the fields of research, etc. And so if the aim is to apply bibliometrics for the comparative evaluation of organizational research performance, the goodness of the algorithms should be assessed on the basis of the precision with which they actually enable measurement of such performance.

To assess it, we compare measurements of the research performance of universities in the Italian academic system, which arise from the application of the previously conformed CvE unsupervised algorithm, with those arising from the use of the supervised algorithm by D'Angelo et al. (2011), hereinafter "DGA." Over more than a decade, this algorithm has been applied by the authors for feeding and continuous updating of the Public Research Observatory of Italy (ORP), a database derived under license from Clarivate Analytics' WoS Core Collection. It indexes the scientific production of Italian academics at individual level, achieving 97% harmonic average of precision and recall (F-measure),[9] thanks to the operation of the DGA algorithm, which avails of a series of "certain" metadata available in the MUR database on university personnel, including their institutional affiliation, academic rank, years of tenure, field of research, and gender (for details see D'Angelo et al., 2011).

Given the maturity of the ORP, developed and refined year by year through the manual correction of the rare false cases, it can be considered a reliable benchmark, or a gold standard, against which to measure the deviations referable to any evaluation conducted using unsupervised methods. The deviations, as we shall see in some detail, are attributable to causes further than simply its lesser abilities in correctly disambiguating authorship. The aim of our work, however, is not to criticize CvE (which is the best alternative when data on staff are not available) but to give bibliometricians, practitioners, and especially decision makers, an idea of the extent of distortions in the research performance ranks of research institutions at overall and area level when forced to use unsupervised algorithms of this kind, rather than supervised ones based on research staff databases, such as DGA.[10]

In a nutshell, a) we evaluate the ability of the rule-based CvE algorithm to disambiguate authorships and the relevant affiliations; then b) we extend the CvE evaluation to its application in research assessment exercises. The paper is organized as follows: Section 2 presents the methodology and describes the data and methods used. In Section 3 we show the results of the analysis. Section 4 concludes, summarizing and also commenting the results, particularly for practitioners and scholars who may wish to replicate the exercise in other geographical and institutional frameworks.

2. DATA AND METHODS

2.1. Identification of Research Staff

The assessment of the comparative research performance of an organization cannot proceed without survey of the scientific activity of its individual researchers, because evaluations that

[8] The affiliation in the byline, in some cases multiple, is not always reliable to unequivocally identify the organization to which the author belongs.

[9] The most frequently used indicators to measure the reliability of bibliometric data sets are precision and recall, which originate from the field of information retrieval (Hjørland, 2010). Precision is the fraction of retrieved instances that are relevant and recall is the fraction of relevant instances that are retrieved.

[10] This is a conservative measure of distortions, as it refers to the application of the best performing unsupervised algorithm according to Tekles and Bornmann (2020).

operate directly at an aggregate level, without accounting for the sectoral distribution of input, produce results with unacceptable error (Abramo & D'Angelo, 2011). Analyses at micro level, however, presuppose precise knowledge of the research staff of the organization, as well as for all "competitor" organizations eligible for comparative evaluation. Adopting an unsupervised approach to solve this task implies using information embedded in bibliometric repositories. For instance, let us consider this publication:

> Abramo, G., & D'Angelo, C. A. (2022). Drivers of academic engagement in public–private research collaboration: An empirical study. *Journal of Technology Transfer, 47*(6), 1861–1884.

WoS supplies, among others, an "address list" field containing, for each author, the relevant affiliation:

- [Abramo, Giovanni] Natl Res Council Italy, Lab Studies Res Evaluat, Inst Syst Anal & Comp Sci IASI CNR, Via Taurini 19, I-00185 Rome, Italy;
- [D'Angelo, Ciriaco Andrea] Univ Rome Tor Vergata Italy, Lab Studies Res Evaluat IASI CNR, Dipartimento Ingn Impresa, Via Politecn 1, I-00133 Rome, Italy.

From this field we can infer that the first author (Giovanni Abramo) is part of the research staff of the National Research Council of Italy and the second one (Ciriaco Andrea D'Angelo), of University of Rome "Tor Vergata."

Aiming at inferring the research staff of all research organizations of a country, one can analyze the set of publications in a given time window showing at least one affiliation of that country. However, it must be taken into account that

- each organization may appear in many different ways (for Roma "Tor Vergata" there occur dozens of variants in 2015–2019 WoS publications)
- the same author may appear under different names in different publications (the second author of the above publication in WoS also appears as Andrea D'Angelo and Andrea Ciriaco D'Angelo)
- many publications list authors with last name and first name initial (D'Angelo, A.; D'Angelo, C.A.; D'Angelo A.C.), and in a very large data set, the cases of homonymy can be numerically very relevant (at University of Rome "Tor Vergata" there are two "D'Angelo, A." professors, and probably just as many nonacademic staff).

All this makes it very complex to disambiguate authors' identity and, consequently, to know "who" works "where."

As explained above, the current study aims to compare the outcomes of the evaluation of research performance by Italian universities, based on two different bibliometric data sets, obtained from the application of two disambiguation methods:

- The first data set, hereinafter "CWTS," relies on the CvE unsupervised approach, a rule-based scoring and oeuvre identification method for disambiguation of authors used for the WoS in-house database of the Centre for Science and Technology Studies (CWTS) at Leiden University.
- The second one hereinafter "ORP," relies on the DGA supervised heuristic approach, which "links" the Italian National Citation Report (indexing all WoS articles by those

authors who indicated Italy as an affiliation country), with data retrieved from the database maintained by the MUR, indexing the full name, academic rank, research field and institutional affiliation of all researchers at Italian universities, at the close of each year.

Much fuller descriptions of the DGA and CvE approaches can be found in D'Angelo and van Eck (2020).

In ORP

- a priori, the availability of MUR data allows precise knowledge of the members of research staff of national universities;
- the census of their scientific production is then carried out by applying the DGA algorithm to the Italian WoS publications.

The CWTS database does not rely on any national research personnel databases to derive the research staff of an organization. It derives it by means of the CvE algorithm. Specifically, to identify the research staff of Italian universities, we apply the CvE algorithm that associates clusters of publications with a cluster label (a kind of "protoindividual") on the basis of the similarity of the publications' metadata. We name the resulting data set of Italian professors the CWTS data set, not to be confused with the CWTS database. In Table 1 we give as an example the information that the algorithm associates with the cluster referred to the second author of the present work (Ciriaco Andrea D'Angelo).

In particular, each protoindividual, uniquely identified by means of a "cluster_id" (56122902 in the example in Table 1), is associated with an "organization" (univ roma tor

Table 1. Description of the output of the CvE approach for one of the authors of this paper

Field	Value
cluster_id	56122902
n_pubs	129
first_year	1996
last_year	2020
"Academic" age	24
full_name	d'angelo, ca
last_name	d'angelo
first_name	ciriaco andrea
email	dangelo@dii.uniroma2.it
organization	univ roma tor vergata
city	rome
country	italy
orcid	0000-0002-6977-6611
researcherid	J-8162-2012

vergata) on the basis of the most recurrent and recent affiliation in the publications assigned to them, as well as an email (dangelo@dii.uniroma2.it) on the basis of the same criterion. For the purposes of our work, we will consider the CvE algorithm as a black box and simply use the output of its application to the Leiden in-house WoS database.

Therefore, the identification of the research staff of each Italian university is based on the "organization" and "email" fields of all clusters identified by CvE and, specifically, all possible variants of the prevailing "organization" and "email" of the protoindividuals indexed in CWTS. As for the "organization", the task has been accomplished by manually labelling all variants in output of the CvE algorithm, imposing country "Italy." As for the "email," the national academic system provides for a standardized web domain (e.g., "uniroma2.it" for Roma "Tor Vergata," "unimi.it" for University of Milano, "unipa.it" for University of Palermo).

Therefore, we can extract relevant information, putting together the two criteria, i.e., concatenating, for each university to be evaluated, two distinct queries. In this regard, the box below shows the query related to the extraction of clusters for University of Rome "Tor Vergata."[11]

```
([organization] = ("state univ rome tor vergata" OR "tor vergata
univ" OR "tor vergata univ rome" OR "univ roma tor vergata" OR "univ
roma tor vergata 2" OR "univ tor vergata")

OR [email] like '%uniroma2.it')

AND [country]='italy'
```

Note that our framework refers to national research assessment exercises, whereby the universities are evaluated on the basis of the performance of the researchers working within them at the time of the launch of the exercise. The underlying rationale is that performance-based research funding looks forward, concentrating on the potential of "current" research staff. Because in Italy professors' mobility is limited (Abramo, D'Angelo, & Di Costa, 2022), we assume that the prevailing "organization" identified by the CvE refers to the current affiliation of the evaluated scholar.

For reasons of robustness, after the first extraction based on the above query, we eliminated universities for which the procedure had identified fewer than 30 clusters (typically telematic or predominantly humanistic universities). For the remaining universities (65 in all), we performed subsequent quality control and tuning operations. The combination of the two conditions ([organization] OR [email]) in fact makes it possible to maximize the recall of the procedure, but also inevitably generates false positives, that is, retrieval of subjects that do not actually belong to the institution in question. Such "incoherent" clusters, include those

- where the "organization" remains ambiguous, or does not refer to a recognized university;

[11] It can be a formidable task, looking at all bibliometric addresses, to identify the variants of "organization" attributable to a single institution (Backes, 2018b). In the present case, we are dealing with 90 universities and manually scanning the 4787 total name variants "associated" with their official emails. Practitioners facing larger numbers or constraints on resources could decide to manually check only the first "n" in terms of frequency. The six "organization" variants in the case shown in the box, for example, account for 95% of all clusters linked with a "%uniroma2.it" email.

- with email not referring to a university;
- with email and organization referring to different universities.

These control operations, for example, lead to the exclusion of the first author of the current article (Giovanni Abramo) from the data set, for whom the initial extraction gives an assignment to the University of Rome "Tor Vergata" because of the email (giovanni.abramo@uniroma2.it), even though his most recurrent organization is "natl res council italy," that is, a nonacademic research body.

Furthermore, to exclude "occasional" and terminated researchers, we impose

- an "academic age" of at least 4 years (given by the difference between the years of the most recent and the first publications);
- the most recent publication of the cluster no earlier than 2020.

Finally, conflicts involving distinct clusters, but sharing the same "ORCID" organization ID or email, are resolved manually.[12]

2.2. Research Performance Measurement

To assess the yearly average performance of each researcher in a period of time, we recur to the Fractional Scientific Strength (FSS_R) indicator of research productivity, defined as

$$FSS_R = \frac{1}{t}\sum_{i=1}^{N}\frac{c_i}{\bar{c}}\mathrm{f_i} \tag{1}$$

where

t = number of years of work of the researcher in the period under observation;[13]
N = number of publications[14] of the researcher in the period under observation;
c_i = citations received by publication i;
$\bar{c}$ = average of distribution of citations received by all publications in the same year and WoS subject category (SC) of publication i;
$\mathrm{f_i}$ = fractional contribution of the researcher to publication i, given by the inverse of the number of coauthors in the byline.

The indicator is calculated over the period 2015–2019, with the citation count at week 13 in 2021. Performance measured at the individual level is then aggregated to obtain the performance of a university (FSS_U) at the SC, area[15] or overall level. In formulae

$$FSS_U = \frac{1}{RS_U}\sum_{j=1}^{RS_U}\frac{FSS_{R_j}}{\overline{FSS_R}} \tag{2}$$

[12] Only 13.3% of the clusters identified through CvE and with affiliation country "Italy" have an ORCID. Therefore, despite the great potential of the ORCID in AD tasks, its use here is only aimed at solving such conflicts and increasing the level of accuracy of the data set.
[13] For researchers in CWTS database we assume $t = 5$ for all.
[14] We consider all publications indexed in the WoS *core collection* (excluding ESCI) with document types: articles, reviews, letters, proceedings.
[15] SCs are classified and grouped into areas according to a system previously published on the webpage of the ISI Journal Citation Reports. This page is no longer available at the current Clarivate site. It should be noted that all SCs are assigned to only one area.

where

RS_U = research staff of the university unit, in the observed period;
FSS_{R_j} = productivity of researcher j in the unit;
$\overline{FSS_R}$ = national average productivity of all productive researchers in the same SC of researcher j.

Prior to any aggregation of data by university unit, it is absolutely necessary that individual performance be scaled against the expected value of the reference SC, but this requires an "SC classification" of each researcher. For this purpose, we used the WoS classification scheme, assigning the prevailing SC as follows:

- for the CWTS-based evaluation, with reference to the researcher's entire scientific production in WoS; in uncertain cases (researcher with multiple prevailing SCs) randomly among those with a higher frequency;
- for the ORP-based evaluation, with reference to the researcher's scientific production in 2001–2019; in uncertain cases (researcher without publications or with multiple prevalent SCs) the one with the highest incidence relative to the SDS[16]-SC pair.

We note that the performance indicator is calculated in the same way for both the ORP and CWTS data sets.[17] Due to its derivation, however, in comparison to the ORP, the CWTS data set has two limitations. First, it does not contain those researchers who have never published in the period under observation, as they cannot be identified from a bibliographic repertory, although they must contribute to the evaluation as they represent a research cost. Second, in the period under observation, the years on staff for each researcher remain unknown, so if one presents publications in CWTS only over 2017–2019, for example, it is unknown whether the lack of production in years 2015–2016 was because they were not on staff. Instead, the information is known in ORP.

For reasons of significance, the analysis excludes researchers in SCs belonging to "Art and Humanities" and "Law, political and social sciences", where the coverage of bibliographic repertories is scarce (Hicks, 1999; Larivière, Archambault et al., 2006; Aksnes & Sivertsen, 2019). The analysis is then restricted to the researchers of SCs pertaining to the STEM and Economics, but also excluding subjects classified in "Multidisciplinary Sciences," and again for the sake of significance, SCs with fewer than 10 observations in both data sets. Table 2 shows the resulting breakdown by area for the two data sets.

Overall, the CWTS-based evaluation concerns 49,908 subjects, 56% more than the 31,989 in the MUR official database, and therefore in ORP. As the ORP data set is the benchmark, we can reasonably consider that, in the CWTS data set, the number of false positives (researchers assigned to a university although not officially part of the research staff) is significantly higher than the number of false negatives (researchers not assigned to a university although part of the research staff).

The data in Table 2 indicate, however, that the overrepresentation of CWTS data set compared to ORP is not identical between the areas. Biomedical research and Clinical medicine

[16] In the Italian university system all professors are classified in one and only one field (named the scientific disciplinary sector, SDS, 370 in all).

[17] This does not mean that the value is necessarily identical for the subjects in both data sets, because a) the CvE and DGA algorithms are not free from error in attributing to a given author the publications they have actually produced; and b) in ORP, the value of "t" may differ from 5.

Table 2. Number of researchers in the two data sets for analysis, by area*

Area	No. of obs data set ORP	No. of obs data set CWTS	Delta
Biology	4928	7987	+62.1%
Biomedical research	2891	7217	+149.6%
Chemistry	1606	2735	+70.3%
Clinical medicine	6205	13444	+116.7%
Earth and space sciences	1937	3039	+56.9%
Economics	3784	1540	−59.3%
Engineering	5554	7873	+41.8%
Mathematics	2055	1805	−12.2%
Physics	2617	3793	+44.9%
Psychology	412	475	+15.3%
Total	31989	49908	+56.0%

* The counts exclude SCs with less than 10 observations in both data sets.

have the largest deviations (+149.6% and +116.7% respectively). In Economics, on the contrary, we have 2.5 times more observations in ORP data set than in CWTS (3784 vs. 1540). A possible reason for such differences is that they might reflect the intensity of academic mobility. Areas characterized by intense mobility are likely to show more false positives, because visiting scholars tend to sign their papers with the hosting university affiliation. However, the level of mobility in Italy is substantially very low and rather homogeneous across fields. On the other hand, the anomaly found for Biomedical research and Clinical medicine could be due to the presence of university hospitals with a large number of nonacademic staff, including physicians and clinical specialists hired under the "national health service" framework and not affiliated to the hosting university.

3. RESULTS

3.1. The Distributions of Performance at the Individual Level

First, we analyze the differences between the two data sets in the distributions of performance at individual level (FSS_R). We report the results of three SCs, exemplary of different types of cases. Table 3 shows the descriptive performance statistics for the authors of "Engineering, manufacturing"; "Dermatology"; "Statistics & probability"; Figure 1 shows the box plots of the distributions for the last two SCs.

For "Engineering, manufacturing," the number of observations in the two data sets results as virtually identical (145 vs. 143), and the distributions seem almost superimposable, with practically identical mean/median and dispersion values. Moreover, the maximum values coincide: referring to the same subject, Professor Francesco Lambiase of the University of L'Aquila. Overall, this accordance occurs in no less than 39 SCs out of the total 160. It can also be noted that in ORP, the minimum value of the distribution is equal to 0, indicating that in the evaluation of the SC using this data set, we find at least one unproductive researcher (FSS = 0), which cannot happen with CWTS (where the minimum recorded is equal to 0.018). Overall,

Table 3. Descriptive statistics of the distribution of research performance (FSS_R) for researchers in three subject categories, comparing ORP and CWTS data sets

		Engineering, manufacturing		Dermatology		Statistics & probability	
		ORP	CWTS	ORP	CWTS	ORP	CWTS
	Obs	145	143	108	258	442	277
	Mean	0.944	1.034	0.942	0.453	0.306	0.353
	Std Dev.	1.102	1.119	1.161	0.818	0.494	0.455
	Variance	1.214	1.253	1.348	0.669	0.244	0.207
	Skewness	3.267	2.958	2.190	3.471	4.887	5.248
	Kurtosis	19.350	16.873	8.791	18.923	36.813	50.275
Percentile	1%	0	0.018	0	0	0	0
	5%	0.023	0.056	0.021	0	0	0.006
	10%	0.056	0.134	0.053	0.004	0	0.028
	25%	0.234	0.357	0.151	0.030	0.049	0.095
	50%	0.683	0.656	0.601	0.139	0.151	0.212
	75%	1.317	1.357	1.203	0.443	0.387	0.454
	90%	2.012	2.342	2.670	1.372	0.749	0.856
	95%	2.733	3.227	3.041	2.162	0.996	1.048
	99%	5.164	4.422	5.499	3.455	2.403	1.666
	Max	8.572	8.572	6.488	6.639	5.034	5.216

this circumstance is detectable in only two SCs besides "Engineering, manufacturing." In contrast, there are 31 SCs where CWTS registers at least one unproductive (FSS = 0), versus ORP finding none.

In "Dermatology," the numerosity of observations in the two data sets is very different, with 258 in CWTS compared to 108 in ORP. In this case the distribution of CWTS performance seems decidedly more shifted to the left than its counterpart in ORP, with lower values both in terms of mean (0.453 vs 0.942) and median (0.139 vs 0.601). In terms of percentiles, the CWTS distribution also shows systematically lower values than the ORP distribution, with the sole exception of the maximum (6.639 vs 6.488).

The situation is diametrically opposed in "Statistics & probability," an SC where the CWTS observations (277) are well below the ORP observations (442). In this case, the CWTS performance distribution appears decisively shifted to the right compared to that for ORP, with higher values of mean (0.353 vs 0.306), median (0.212 vs 0.151), and all percentiles.

Figures 2 and 3 show the comparison of the mean and median values of the distributions for all 160 SCs, in the two data sets. At a glance, one can observe a greater number of cases in which the "central" values (mean and median) calculated in CWTS are lower than their counterparts recorded in ORP. Indeed, in 107 SCs out of 160 (i.e., 67%), the mean value of the FSS_R is higher in the ORP relative to the CWTS data set, and for the medians, this occurs in 114 SCs out of 160 (71%).

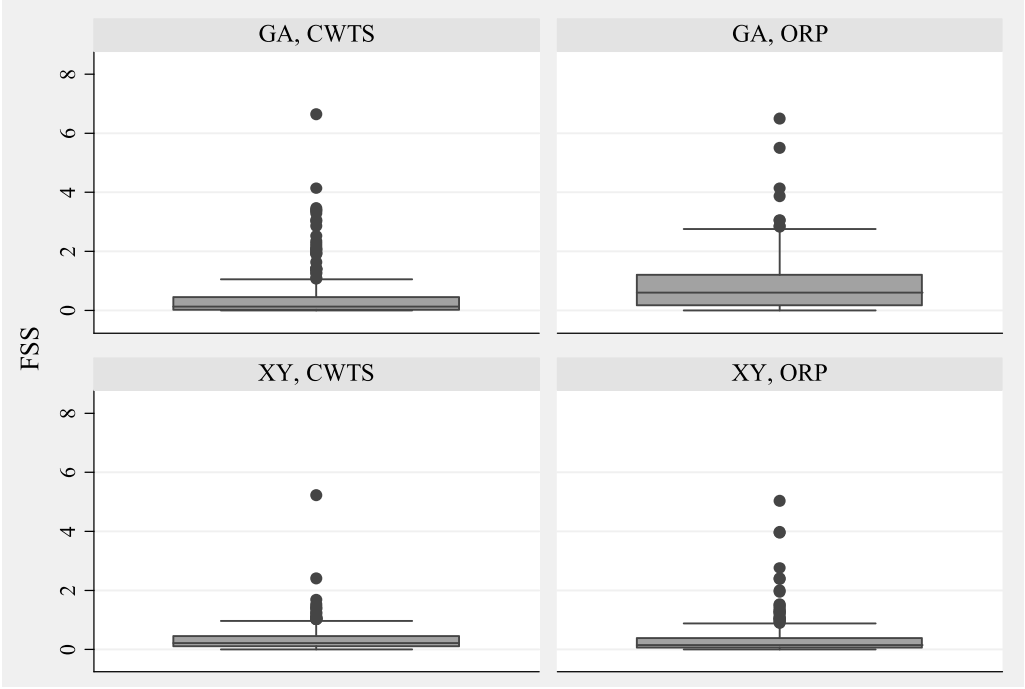

Figure 1. Boxplots of research performance (FSS_R) distribution for researchers in Dermatology (GA) and Statistics & probability (XY), comparing ORP and CWTS data sets.

Thus, in general, it would appear that the distribution of performance recorded in the ORP data set is more rightward shifted than in the CWTS data set. This could be explained by the different composition of the two data sets and, in particular, the overrepresentation of CWTS compared to ORP. In fact, the Pearson ρ correlation between the deviations in terms of the number of observations and the deviations between the mean FSS_R values for the 160 SCs is −0.515 (−0.454 when considering the medians). Basically, in the SCs where the overrepresentation of the CWTS data set compared to ORP is greater, the mean performance values in CWTS are significantly lower than those found in ORP, and vice versa. We can hypothesize

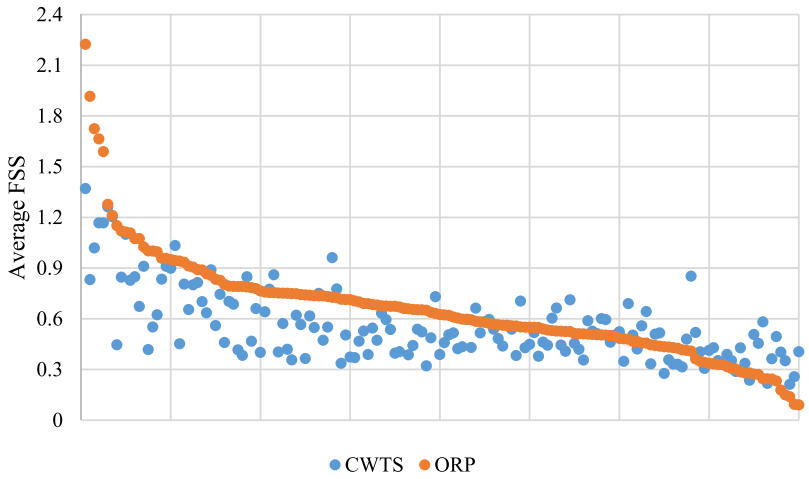

Figure 2. Distribution of the mean values of FSS_R detected in the two data sets, for the 160 subject categories considered.

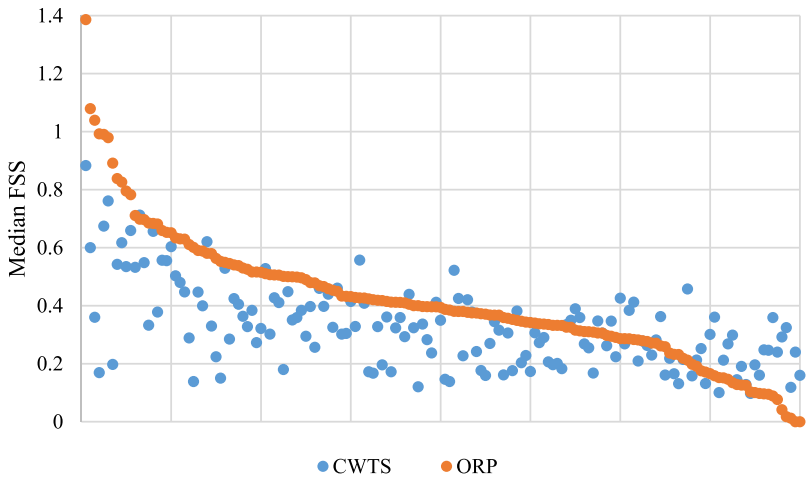

Figure 3. Distribution of the median FSS_R values detected in the two data sets, for the 160 subject categories considered.

that the so-called *false positives* have lower average performance values than the "true positives." In other words, "nonfaculty" personnel, but with a bibliometrically prevalent university affiliation, have a lower average FSS_R than the research staff truly on faculty at the universities in the same field of observation. This has an important implication concerning the use of CWTS data sets for comparative performance assessment, in that it would evidently "penalize" those organizations (and areas within organizations) with a higher concentration of nonfaculty personnel in their research staff.

3.2. Evaluation of the Universities' Performance

We now move on to analyze the deviations between the two data sets in terms of the score and rank of the performance of the universities, through the FSS_U indicator in formula [2]. Figure 4 shows the values measured at the overall level for the two data sets of the 65 universities with at least 30 observations of researchers. The dispersion of the values for the ORP data set is greater than that for the CWTS (standard deviations 0.316 vs 0.175). Figure 5 instead shows

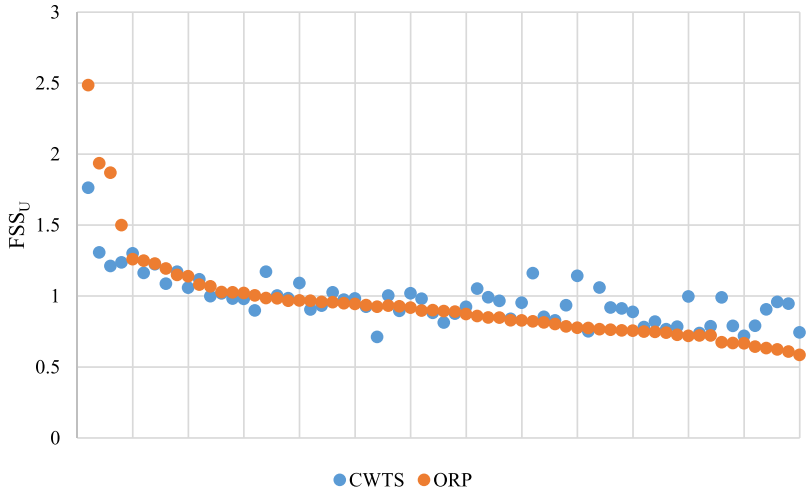

Figure 4. Distribution of FSS_U for universities in the CWTS and ORP data sets, at the overall level.

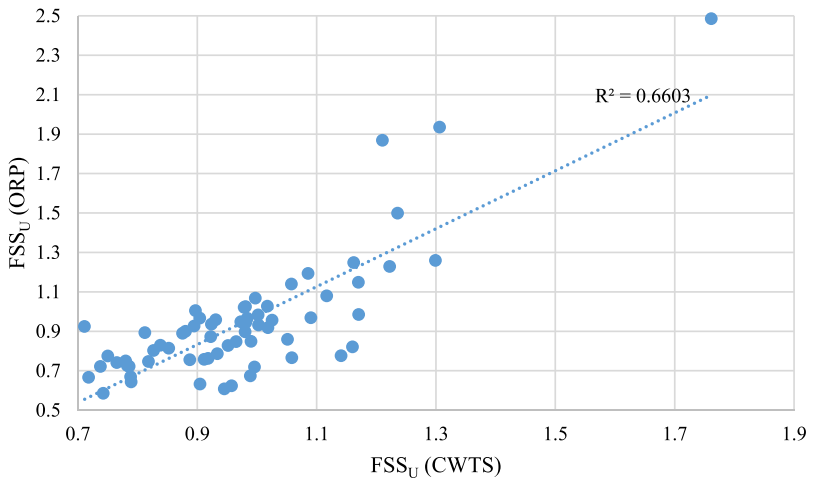

Figure 5. Scatterplot of FSS_U values for universities in the CWTS and ORP data sets, at the overall level.

the scatterplot of the values of the two indicators and evidences a strong correlation. The detailed values are shown in Table 4: the University Vita-Salute San Raffaele and the *Scuola Superiore* S. Anna are at the top in both assessments. The top 11 universities between the two rankings vary by at most six positions (the case for the University of Padua and Politecnico di Milano). Thus, the evaluation by the CWTS data set returns a top part of the ranking substantially similar to that resulting in ORP. Instead, the situation in the middle and lower part of the ranking is different, where LUISS stands out, which shifts 36 positions between the two rankings, and the University "Campus Bio-medico," which shifts 29. On the other hand, there are also five universities (University of Naples "Parthenope"; University of Enna; University of the Mediterranean Studies of Reggio Calabria; University of Sannio; University of Teramo) which in the CWTS-based ranking gain between 31 and 34 positions compared to ORP. It is noticeable the gain in performance score by the bottom-ranked universities in the CWTS-based ranking. A possible interpretation is that the ORP performance of true positives is so low that false positives cannot help contributing to increase overall CWTS-based performance.

The magnitude and direction of these variations could be related to the differing numbers of researchers evaluated in the two modes. In fact, the percentage deviations of FSS_U between the two data sets are significantly and negatively correlated with the percentage deviations of numerosity (Pearson $\rho = -0.526$). The same is true for the ranking jumps and the percentage deviations in numerosity between the two data sets (Pearson $\rho = -0.361$).

To summarize, compared to the ORP benchmark, the value of the performance recorded in CWTS (both absolute and relative) decreases as the concentration of "nonfaculty" personnel in the research staff increases. Thus, the hypothesis is confirmed that conducted via CWTS, universities (and areas within universities) with a higher proportion of nonfaculty research staff are actually penalized. This does not prevent the Vita-Salute San Raffaele University from coming out on top in both rankings, despite the fact that in the CWTS data set there are 364 researchers associated with it compared to the 103 actual researchers recorded in ORP.

Depending on the particular application one has in mind (e.g., for a policymaker, or a program administrator), the sensitivity of any measuring instrument will be more or less critical. In the case of the data in Table 4, a variation to the second decimal place in the values of FSS_U results in some major jumps in ranking. This certainly prompts thinking that a less precise ranking would be reasonable, such as by performance classes, such as by quartiles, which are in

Table 4. Performance score and rank of Italian universities, comparing CWTS and ORP data sets

	CWTS				ORP				
University	Obs	FSS_U	Rank	Perc.	Obs	FSS_U	Rank	Perc.	Δ Rank
Vita-Salute San Raffaele	364	1.762	1	100	103	2.485	1	100	0
Scuola Superiore S.Anna	172	1.306	2	98	82	1.935	2	98	0
SISSA	128	1.210	6	92	67	1.868	3	97	−3
Libera Università di Bolzano	103	1.236	4	95	91	1.499	4	95	0
Commerciale Luigi Bocconi	112	1.299	3	97	191	1.259	5	94	+2
Politecnico di Bari	264	1.163	9	88	212	1.248	6	92	−3
Trento	487	1.223	5	94	332	1.228	7	91	+2
Padova	2845	1.086	14	80	1394	1.193	8	89	−6
Salerno	696	1.170	8	89	523	1.148	9	88	+1
Politecnico di Milano	1339	1.058	16	77	993	1.139	10	86	−6
Napoli "Federico II"	2405	1.117	12	83	1659	1.079	11	84	−1
Milano	2724	0.998	23	66	1317	1.068	12	83	−11
Pisa	1622	1.018	20	70	908	1.027	13	81	−7
Verona	838	0.981	29	56	421	1.025	14	80	−15
Firenze	1959	0.979	31	53	983	1.020	15	78	−16
"Campus Bio-medico"	257	0.897	45	31	107	1.004	16	77	−29
Istituto Univ. di Scienze Motorie	50	1.171	7	91	42	0.985	17	75	+10
Perugia	968	1.002	22	67	654	0.983	18	73	−4
Catania	976	1.091	13	81	708	0.968	19	72	+6
Torino	2166	0.904	44	33	1121	0.966	20	70	−24
Politecnico di Torino	988	0.984	27	59	658	0.966	21	69	−6
Ferrara	700	0.931	38	42	389	0.958	22	67	−16
Magna Grecia di Catanzaro	290	1.026	18	73	162	0.956	23	66	+5
Pavia	957	0.973	32	52	536	0.948	24	64	−8
Bologna	2889	0.982	28	58	1640	0.944	25	63	−3
Politecnica delle Marche	707	0.924	39	41	436	0.936	26	61	−13
Tuscia	254	1.003	21	69	177	0.932	27	59	+6
Bergamo	132	0.895	46	30	145	0.926	28	58	−18
LUISS	32	0.711	65	0	51	0.924	29	56	−36
Calabria	627	1.019	19	72	473	0.918	30	55	+11
Cattolica del Sacro Cuore	906	0.881	48	27	704	0.900	31	53	−17

Table 4. (*continued*)

University	CWTS				ORP				Δ Rank
	Obs	FSS_U	Rank	Perc.	Obs	FSS_U	Rank	Perc.	
Milano Bicocca	1020	0.980	30	55	567	0.897	32	52	+2
Urbino "Carlo Bo"	211	0.813	54	17	159	0.893	33	50	−21
del Salento	382	0.876	49	25	298	0.889	34	48	−15
Insubria	360	0.923	40	39	241	0.872	35	47	−5
Messina	810	1.052	17	75	642	0.859	36	45	+19
Roma Tre	397	0.990	25	63	349	0.848	37	44	+12
Foggia	275	0.965	33	50	211	0.847	38	42	+5
Genova	1257	0.839	51	22	745	0.828	39	41	−12
dell'Aquila	495	0.952	35	47	383	0.827	40	39	+5
Napoli "Parthenope"	213	1.160	10	86	223	0.820	41	38	+31
Roma "La Sapienza"	3414	0.852	50	23	2048	0.813	42	36	−8
Brescia	702	0.827	52	20	410	0.802	43	34	−9
Roma "Tor Vergata"	1158	0.934	37	44	849	0.785	44	33	+7
Enna	38	1.141	11	84	56	0.776	45	31	+34
Cagliari	798	0.750	61	6	560	0.774	46	30	−15
Mediterranea di Reggio Calabria	181	1.059	15	78	159	0.765	47	28	+32
Gabriele D'Annunzio	587	0.918	41	38	394	0.762	48	27	+7
Palermo	1181	0.912	42	36	885	0.757	49	25	+7
Bari	1193	0.888	47	28	852	0.755	50	23	+3
Parma	895	0.780	59	9	610	0.749	51	22	−8
Trieste	536	0.818	53	19	378	0.746	52	20	−1
Camerino	288	0.765	60	8	200	0.741	53	19	−7
Modena e Reggio Emilia	888	0.783	58	11	537	0.726	54	17	−4
Siena	752	0.786	57	13	392	0.722	55	16	−2
Ca' Foscari Venezia	198	0.738	63	3	184	0.721	56	14	−7
Sannio	154	0.996	24	64	142	0.719	57	13	+33
Teramo	115	0.989	26	61	104	0.674	58	11	+32
Udine	526	0.789	56	14	402	0.668	59	9	+3
Piemonte Orientale A. Avogadro	178	0.718	64	2	234	0.666	60	8	−4
Sassari	474	0.790	55	16	342	0.643	61	6	+6
Molise	166	0.905	43	34	133	0.632	62	5	+19

Table 4. (*continued*)

	CWTS				ORP				
University	Obs	FSS_U	Rank	Perc.	Obs	FSS_U	Rank	Perc.	Δ Rank
Seconda Napoli	574	0.958	34	48	563	0.623	63	3	+29
Cassino	146	0.946	36	45	152	0.608	64	2	+28
Basilicata	257	0.743	62	5	222	0.585	65	0	+3

fact used in some national research evaluation exercises. In Table 5 we report the ranking of the universities' performance in the two data sets, by quartiles. The data show that 36 out of 65 universities are ranked in the same quartile in the two modes (in the main diagonal). The remaining 29 show a jump of at least one quartile. Of these

- 13 have a better ranking based on CWTS than ORP (above the main diagonal);
- 16 have a worse ranking based on CWTS than ORP (below the main diagonal).

The nine universities shown in Table 6 experience a two-quartile jump between the two rankings. No university experiences a three-quartile jump, that is, top to bottom or vice versa.

What was just seen in Section 3.2 at the overall level is repeated at the area level. Figure 6 shows the scatterplot of FSS_U calculated in the two modes, at this level of aggregation. Table 7

Table 5. Performance quartiles of the universities evaluated using CWTS and ORP data sets

		ORP			
		I	II	III	IV
CWTS	I	12	1	4	0
	II	4	8	2	2
	III	1	5	6	4
	IV	0	2	4	10

Table 6. Universities showing two-quartile ranking jumps between the two data sets

Ateneo	Quartile CWTS	Quartile ORP
Messina	1	3
Napoli "Parthenope"	1	3
Enna	1	3
Mediterranea di Reggio Calabria	1	3
Sannio	2	4
Teramo	2	4
"Campus Bio-medico"	3	1
LUISS	4	2
Urbino "Carlo Bo"	4	2

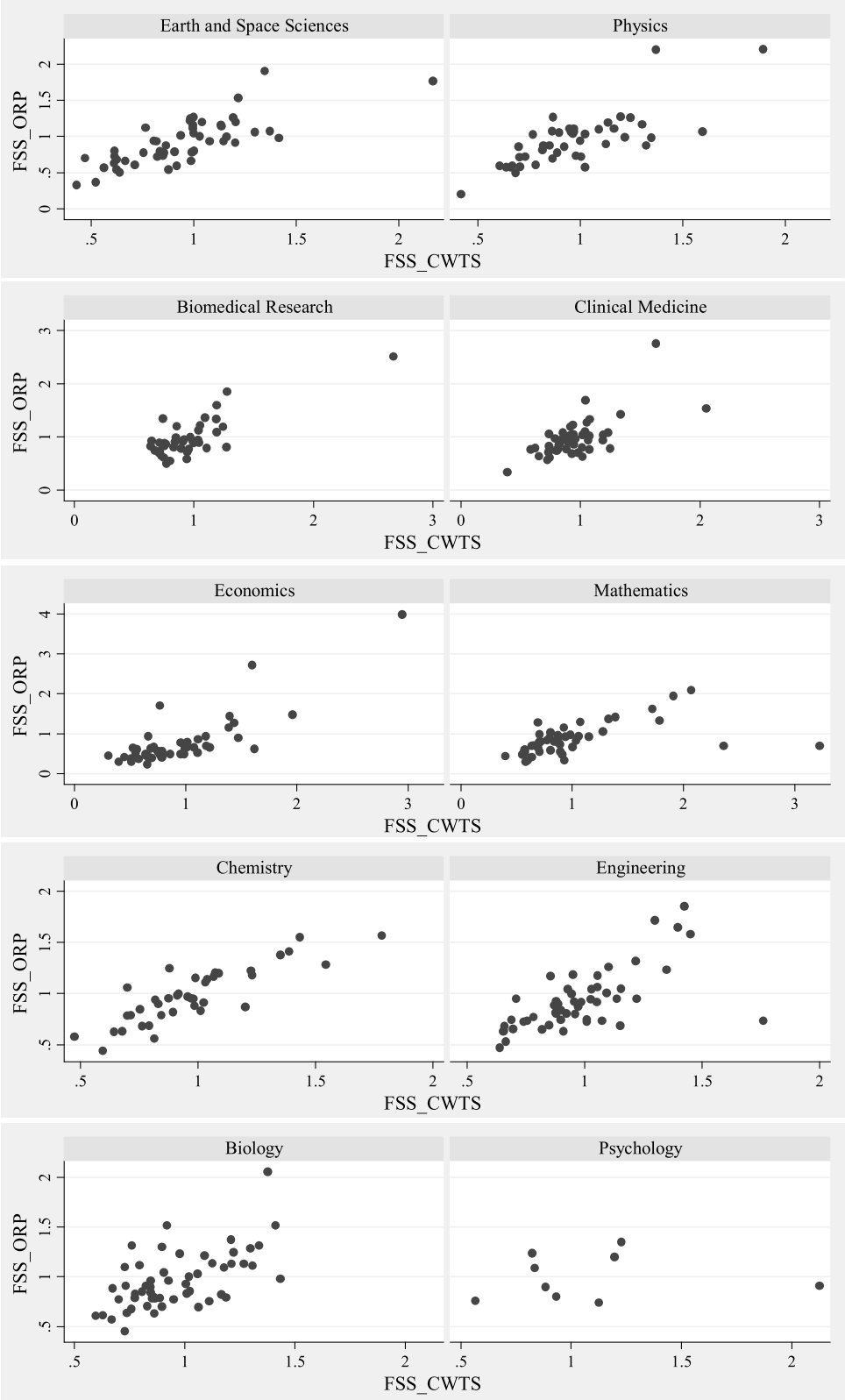

Figure 6. Distributions of performance evaluated in the two modes, by area.

Table 7. Correlation between score and ranking of the 65 Italian universities, evaluated in the two modes, by area

Discipline	No. of obs (evaluated universities)	Pearson ρ	Spearman ρ
Biology	54	0.602	0.576
Biomedical Research	42	0.789	0.537
Chemistry	39	0.851	0.824
Clinical Medicine	47	0.681	0.580
Earth and Space Sciences	50	0.775	0.747
Economics	48	0.808	0.712
Engineering	52	0.657	0.673
Mathematics	48	0.504	0.607
Physics	44	0.760	0.702
Psychology	9	0.072	0.233
Overall	65	0.813	0.686

presents the correlation between score and ranking of the 65 universities evaluated by area. At the score level, the Pearson coefficient recorded in the two assessments at the overall level is 0.816. The figure for the individual areas is never less than 0.5 with the sole exception of Psychology, which is an area with a very low number of assessable universities (only nine). The correlation coefficients between the ranks are slightly lower, that is, 0.686 at the overall level. This confirms the exception of Psychology, an area in which the two assessment modes lead to results that are completely noncorrelated. Chemistry is confirmed as the area with the most closely aligned ranks between the two modes (Spearman 0.830) followed by Earth and Space Sciences (0.747) and Economics (0.712).

4. DISCUSSION AND CONCLUSIONS

Evaluative bibliometricians have been engaged for years in the continuous improvement of indicators and methods for evaluating the scientific activities of individuals, organizations, and national and territorial research systems. The major obstacle hindering a leap ahead in evaluation techniques is the lack of input data. Yet a basic principle is that the proper evaluation of the performance of any subject at any level, including in research performance requires, in addition to output, input data. In particular, assessing the performance of universities requires knowledge of the individual researchers working within them (Abramo & D'Angelo, 2016a, 2016b). In large-scale research assessments, the evaluators have at least three different options to acquire input (and output) data:

- They can ask the institutions being evaluated a direct involvement in declaring and submitting their research staff (as well as research products);
- They may draw a list of unique identifiers for institutions and/or authors (and then use these for querying a bibliometric database for having their research products);
- They can extract publications from a bibliometric repertoire and, then, disambiguate the true identity of the relevant authors and their institutions.

These approaches present significant trade-offs: the first one can guarantee a high level of precision and recall but is particularly "costly" because of the opportunity cost of the surveyed subjects for collecting and selecting inputs and outputs for the evaluation.

The introduction of unique identifiers for researchers and organizations (ROR, ORCID, etc.) is important and necessary for improving the quality of research information systems (Enserink, 2009), but at the moment the coverage is limited and not uniform in terms of country and/or field (Youtie, Carley et al., 2017).

The third option implies setting up a large-scale bibliometric database in "desk mode" and offers rapid and economical implementation. However, the task is challenging because of homonyms in author names and variations in the way authors indicate their name and affiliation. Such challenges have determined the development and continuous improvement of disambiguation methods.

In this work, we have focused on the issue of the reliability of author disambiguation algorithms in identifying the true publications of each observed subject in combination with their ability to identify the research staff of the home institutions, including their placement in disciplinary fields.

In particular, we evaluated the goodness-of-fit of the CvE unsupervised author-name disambiguation algorithm in measuring the performance scores and ranks of Italian universities, operating through direct processing of bibliometric data for deduction of the research staff (and thus the input data) of each university. The validation was carried out through the comparison with the DGA algorithm, based on the a priori knowledge of the research staff officially in post in each national university.

The results of the comparison showed that the application of the unsupervised approach leads to an overestimation of the research staff of an organization. Overall, for the field of observation adopted in this study, this meant 56% more subjects in the CWTS data set than in the ORP data set, which draws on guaranteed data from the MUR. One of the reasons for this would be that the CvE approach, which underlies the CWTS data set, attributes all researchers to an organization when these have prevalently indicated the relative affiliation in signing their scientific publications, independently of their effective position within the organization. In this way, doctoral and postdoctoral students, postdoctoral fellows, visiting scholars, collaborators, and a range of other individuals who would not be eligible for evaluation in an official national evaluation exercise also end up on a university's list of researchers.[18]

It should also be considered that the CvE algorithm[19] tends to favour precision over recall; in particular, the publication oeuvre of an author can be split over multiple "clusters" if not enough proof is found for joining publications together. This means that the actual value of the over-representation of an organization's research staff in the CWTS data set is lower than the figure measurable by direct comparison with the ORP data. At an overall level, therefore, that +56% represents an upper bound of the actual incidence of nonfaculty personnel in the CWTS data set.

Having said this, the scores and ranks recorded in the two compared modes show a significant and rather high correlation: At the overall level, Pearson and Spearman coefficients, respectively, are 0.813 and 0.686. At the area level, the values are never below 0.5 with peaks in Chemistry (0.851; 0.824). The only critical area is Psychology, however this is an area present only in a small

[18] Although including doctoral and post doctoral students or postdoctoral fellows may be problematic from the viewpoint of the Italian research assessment system, it may not be necessarily so in other research assessment systems.

[19] For convenience, we refer here to the CvE algorithm, but our conclusions refer to any unsupervised approaches to author disambiguation.

number of assessed universities, at nine. Still, the overall correlation covers significant jumps at local level: comparing the CWTS-based assessment to the benchmark, in the ranking of all 69 assessed universities, nine show deviations of two quartiles, better or worse.

Among the drivers of these deviations, certainly the greatest weight goes to the different number of observations between the two data sets. Empirically, it emerges that percentage deviations in FSS for universities are significantly and negatively correlated with percentage deviations in numerosity between the two data sets.

If we assume that the overrepresentation of the CWTS data set with respect to ORP depends largely on the CWTS inclusion of "nonfaculty" personnel, we can deduce that on average, these personnel perform less well. More importantly, it can also be concluded that an evaluation conducted by means of the CvE approach, although there could be exceptions, would generally penalize universities (and areas) with a higher proportion of nonfaculty research-active personnel. A comparison of the research performance of "nonfaculty" personnel vs. "faculty" personnel could be the object of future research.

Obviously this effect, which is certainly significant, must be discounted against the intrinsic limits of CvE. Notably, as mentioned above, such an algorithm can in some cases attribute a researcher's scientific production to two (or more) distinct "clusters," especially in the presence of a scientific production characterized by heterogeneous and highly differentiated bibliometric metadata. The impact of such "splitting" on the outcomes of the comparative evaluation of universities should, however, be very limited, as the splitting cases should be evenly distributed and not focus on researchers from one organization over those from others. And, it is worth remembering, the literature in any case indicates CvE as the best performing of such unsupervised algorithms (Tekles & Bornmann, 2020).

While waiting for policymakers to take action towards national and international systems for collecting input data, which would enable bibliometricians to carry out what the same policymakers are ever more insistently demanding, practitioners may consider using the CWTS data as in this current paper. In particular, the methodology described here makes it possible for others to replicate the comparative analyses in the frameworks of their interest (national or international), simply by processing the output of the CvE algorithm in an appropriate manner, in particular by considering the relative institutions' official URL domains. A notable side benefit of this would be that with this, practitioners now have a precise measure of the extent of distortions inherent in any evaluation exercises using unsupervised algorithms. And, for policymakers, knowing the extent of performance measure distortions reveals useful in deciding whether to invest in development of databases of national research personnel, or to settle for the less precise assessments.

ACKNOWLEDGMENTS

We are indebted to the Centre for Science and Technology Studies (CWTS) at Leiden University for providing us with access to the in-house WoS database from which we extracted data at the basis of our elaborations.

AUTHOR CONTRIBUTIONS

Giovanni Abramo: Conceptualization; Investigation; Methodology; Supervision; Writing—Original draft; Writing—Review & editing. Ciriaco Andrea D'Angelo: Conceptualization; Data curation; Investigation; Methodology; Visualization; Writing—Original draft; Writing—Review & editing.

COMPETING INTERESTS

The authors have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

FUNDING INFORMATION

The research project received no funding by third parties.

DATA AVAILABILITY

The bibliometric data set used in this study has been extracted from the CWTS in-house WoS database, made available under license by Clarivate Analytics. The authors are not allowed to redistribute WoS data.

REFERENCES

Abramo, G., & D'Angelo, C. A. (2011). National-scale research performance assessment at the individual level. *Scientometrics*, *86*(2), 347–364. https://doi.org/10.1007/s11192-010-0297-2

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2011). A national-scale cross-time analysis of university research performance. *Scientometrics*, *87*(2), 399–413. https://doi.org/10.1007/s11192-010-0319-0

Abramo, G., & D'Angelo, C. A. (2016a). A farewell to the MNCS and like size-independent indicators. *Journal of Informetrics*, *10*(2), 646–651. https://doi.org/10.1016/j.joi.2016.04.006

Abramo, G., & D'Angelo, C. A. (2016b). A farewell to the MNCS and like size-independent indicators: Rejoinder. *Journal of Informetrics*, *10*(2), 679–683. https://doi.org/10.1016/j.joi.2016.01.011

Abramo, G., & D'Angelo, C. A. (2016c). A comparison of university performance scores and ranks by MNCS and FSS. *Journal of Informetrics*, *10*(4), 889–901. https://doi.org/10.1016/j.joi.2016.07.004

Abramo, G., Aksnes, D. W., & D'Angelo, C. A. (2020). Comparison of research productivity of Italian and Norwegian professors and universities. *Journal of Informetrics*, *14*(2), 101023. https://doi.org/10.1016/j.joi.2020.101023

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2022). The effect of academic mobility on research performance: The case of Italy. *Quantitative Science Studies*, *3*(2), 345–362. https://doi.org/10.1162/qss_a_00192

Abramo, G., D'Angelo, C. A., & Reale, E. (2019). Peer review vs bibliometrics: Which method better predicts the scholarly impact of publications? *Scientometrics*, *121*(1), 537–554. https://doi.org/10.1007/s11192-019-03184-y

Abramo, G., D'Angelo, C. A., & Rosati, F. (2013). Measuring institutional research productivity for the life sciences: The importance of accounting for the order of authors in the byline. *Scientometrics*, *97*(3), 779–795. https://doi.org/10.1007/s11192-013-1013-9

Aksnes, D. W., Schneider, J. W., & Gunnarsson, M. (2012). Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalised counting methods. *Journal of Informetrics*, *6*(1), 36–43. https://doi.org/10.1016/j.joi.2011.08.002

Aksnes, D. W., & Sivertsen, G. (2019). A criteria-based assessment of the coverage of Scopus and Web of Science. *Journal of Data and Information Science*, *4*(1), 1–21. https://doi.org/10.2478/jdis-2019-0001

Backes, T. (2018a). Effective unsupervised author disambiguation with relative frequencies. In J. Chen, M. A. Gonçalves, & J. M. Allen (Eds.), *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (pp. 203–212). Fort Worth, TX: Association for Computing Machinery. https://doi.org/10.1145/3197026.3197036

Backes, T. (2018b). The impact of name-matching and blocking on author disambiguation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 803–812). Turin, Italy: Association for Computing Machinery. https://doi.org/10.1145/3269206.3271699

Butler, D. (2010). University rankings smarten up. *Nature*, *464*(7285), 16–17. https://doi.org/10.1038/464016a, PubMed: 20203575

Caron, E., & van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In E. Noyons (Ed.), *Proceedings of the Science and Technology Indicators Conference 2014 Leiden* (pp. 79–86). Leiden: Universiteit Leiden—CWTS.

Cota, R. G., Gonçalves, M. A., & Laender, A. H. F. (2007). A heuristic-based hierarchical clustering method for author name disambiguation in digital libraries. Paper presented at the XXII Simpósio Brasileiro de Banco de Dados, João Pessoa.

D'Angelo, C. A., & Abramo, G. (2015). Publication rates in 192 research fields. In A. Salah, Y. Tonta, A. A. A. Salah, C. Sugimoto (Eds.), *Proceedings of the 15th International Society of Scientometrics and Informetrics Conference - (ISSI 2015)* (pp. 909–919). Istanbul: Bogazici University Printhouse.

D'Angelo, C. A., & van Eck, N. J. (2020). Collecting large-scale publication data at the level of individual researchers: A practical proposal for author name disambiguation. *Scientometrics*, *123*(2), 883–907. https://doi.org/10.1007/s11192-020-03410-y

D'Angelo, C. A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. *Journal of the American Society for Information Science and Technology*, *62*(2), 257–269. https://doi.org/10.1002/asi.21460

Dehon, C., McCathie, A., & Verardi, V. (2010). Uncovering excellence in academic rankings: A closer look at the Shanghai ranking. *Scientometrics*, *83*(2), 515–524. https://doi.org/10.1007/s11192-009-0076-0

Enserink, M. (2009). Are you ready to become a number? *Science*, *323*(5922), 1662–1664. https://doi.org/10.1126/science.323.5922.1662, PubMed: 19325094

Gauffriau, M., & Larsen, P. O. (2005). Counting methods are decisive for rankings based on publication and citation studies. *Scientometrics*, *64*(1), 85–93. https://doi.org/10.1007/s11192-005-0239-6

Gläser, J., & Laudel, G. (2016). Governing science: How science policy shapes research content. *Archives Europeennes De Sociologie*, *57*(1), 117–168. https://doi.org/10.1017/S0003975616000047

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, *44*(2), 193–215. https://doi.org/10.1007/BF02457380

Hjørland, B. (2010). The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technology*, *61*(2), 217–237. https://doi.org/10.1002/asi.21261

Huang, M. H., Lin, C. S., & Chen, D. Z. (2011). Counting methods, country rank changes, and counting inflation in the assessment of national research productivity and impact. *Journal of the American Society for Information Science and Technology*, *62*(12), 2427–2436. https://doi.org/10.1002/asi.21625

Hussain, I., & Asghar, S. (2018). DISC: Disambiguating homonyms using graph structural clustering. *Journal of Information Science*, *44*(6), 830–847. https://doi.org/10.1177/0165551518761011

Iglesias, J. E., & Pecharromán, C. (2007). Scaling the h-index for different scientific ISI fields. *Scientometrics*, *73*(3), 303–320. https://doi.org/10.1007/s11192-007-1805-x

Karlsson, S. (2017). Evaluation as a travelling idea: Assessing the consequences of research assessment exercises. *Research Evaluation*, *26*(2), 55–65. https://doi.org/10.1093/reseval/rvx001

Larivière, V., Archambault, É., Gingras, Y., & Vignola-Gagné, É. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, *57*(8), 997–1004. https://doi.org/10.1002/asi.20349

Liu, W., Doğan, R. I., Kim, S., Comeau, D. C., Kim, W., Yeganova, L., ... Wilbur, W. J. (2014). Author name disambiguation for PubMed. *Journal of the Association for Information Science and Technology*, *65*(4), 765–781. https://doi.org/10.1002/asi.23063, PubMed: 28758138

Lillquist, E., & Green, S. (2010). The discipline dependence of citation statistics. *Scientometrics*, *84*(3), 749–762. https://doi.org/10.1007/s11192-010-0162-3

Moed, H. F. (2010). CWTS crown indicator measures citation impact of a research group's publication oeuvre. *Journal of Informetrics*, *4*(3), 436–438. https://doi.org/10.1016/j.joi.2010.03.009

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. https://doi.org/10.1002/asi.22746

Rinia, E. J., De Lange, C., & Moed, H. F. (1993). Measuring national output in physics: Delimitation problems. *Scientometrics*, *28*(1), 89–110. https://doi.org/10.1007/BF02016287

Rose, M. E., & Kitchin, J. R. (2019). pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. *SoftwareX*, *10*, 100263. https://doi.org/10.1016/j.softx.2019.100263

Sandström, U., & Sandström, E. (2009). Meeting the micro-level challenges: Bibliometrics at the individual level. In *12th International Conference on Scientometrics and Informetrics* (pp. 845–856). Rio de Janeiro, Brazil.

Schulz, C., Mazloumian, A., Petersen, A. M., Penner, O., & Helbing, D. (2014). Exploiting citation networks for large-scale author name disambiguation. *EPJ Data Science*, *3*, 11. https://doi.org/10.1140/epjds/s13688-014-0011-3

Sorzano, C. O. S., Vargas, J., Caffarena-Fernández, G., & Iriarte, A. (2014). Comparing scientific performance among equals. *Scientometrics*, *101*(3), 1731–1745. https://doi.org/10.1007/s11192-014-1368-6

Tekles, A., & Bornmann, L. (2020). Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. *Quantitative Science Studies*, *1*(4), 1510–1528. https://doi.org/10.1162/qss_a_00081

Turner, D. (2005). Benchmarking in universities: League tables revisited. *Oxford Review of Education*, *31*(3), 353–371. https://doi.org/10.1080/03054980500221975

van Hooydonk, G. (1997). Fractional counting of multi-authored publications: Consequences for the impact of authors. *Journal of the American Society for Information Science*, *48*(10), 944–945. https://doi.org/10.1002/(SICI)1097-4571(199710)48:10<944::AID-ASI8>3.0.CO;2-1

van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, *62*(1), 133–143. https://doi.org/10.1007/s11192-005-0008-6

Waltman, L., & van Eck, N. J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of Informetrics*, *9*(4), 872–894. https://doi.org/10.1016/j.joi.2015.08.001

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37–47. https://doi.org/10.1016/j.joi.2010.08.001

Wu, H., Li, B., Pei, Y., & He, J. (2014). Unsupervised author disambiguation using Dempster–Shafer theory. *Scientometrics*, *101*(3), 1955–1972. https://doi.org/10.1007/s11192-014-1283-x

Wu, J., & Ding, X.-H. (2013). Author name disambiguation in scientific collaboration and mobility cases. *Scientometrics*, *96*(3), 683–697. https://doi.org/10.1007/s11192-013-0978-8

Youtie, J., Carley, S., Porter, A. L., & Shapira, P. (2017). Tracking researchers and their outputs: New insights from ORCIDs. *Scientometrics*, *113*(1), 437–453. https://doi.org/10.1007/s11192-017-2473-0

Zacharewicz, T., Lepori, B., Reale, E., & Jonkers, K. (2019). Performance-based research funding in EU member states—A comparative assessment. *Science and Public Policy*, *46*(1), 105–115. https://doi.org/10.1093/scipol/scy041

Zhu, J., Wu, X., Lin, X., Huang, C., Fung, G. P. C., & Tang, Y. (2017). A novel multiple layers name disambiguation framework for digital libraries using dynamic clustering. *Scientometrics*, *114*(3), 781–794. https://doi.org/10.1007/s11192-017-2611-8

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics*, *63*(2), 373–401. https://doi.org/10.1007/s11192-005-0218-y