Journal of Informetrics 10 (2016) 1055–1078

Contents lists available at ScienceDirect

Journal of Informetrics

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

A critical cluster analysis of 44 indicators of author-level performance

Lorna Wildgaard

Royal School of Library and Information Science, Faculty of the Humanities, Copenhagen University, Birketinget 6, 2300 Copenhagen S, Denmark

ARTICLE INFO

Article history:
Received 28 April 2016
Received in revised form
20 September 2016
Accepted 20 September 2016
Available online 5 October 2016

ABSTRACT

This paper explores a 7-stage cluster methodology as a process to identify appropriate indicators for evaluation of individual researchers at a disciplinary and seniority level. Publication and citation data for 741 researchers from 4 disciplines was collected in Web of Science. Forty-four indicators of individual researcher performance were computed using the data. The clustering solution was supported by continued reference to the researcher's curriculum vitae, an effect analysis and a risk analysis. Disciplinary appropriate indicators were identified and used to divide the researchers into four groups; low, middle, high and extremely high performers. Seniority-specific indicators were not identified. The practical importance of the recommended disciplinary appropriate indicators is concerning. Our study revealed several critical concerns that should be investigated in the application of statistics in research evaluation.

The strength of the 7-stage cluster methodology is that it makes clear that in the evaluation of individual researchers, statistics cannot stand alone. The methodology is reliant on contextual information to verify the bibliometric values and cluster solution. It is important to do studies that investigate the usefulness of statistical evaluation methodologies to help us as a community learn more about the appropriateness of particular bibliometric indicators in the analysis of different researcher profiles.

1. Introduction

"Quality nowadays seems to a large extent to be defined as productivity," wrote Arensbergen in 2014. Researchers are defined and are defining themselves in assessments in terms of their performance on bibliometric indicators of production and citation impact (Martin & Irvine, 1983). Developing indicators that most accurately capture the researcher's performance has led to an explosion in the generation of author-level indicators, (Wildgaard, 2015a; Iliev, 2014). Yet even though there is now a multitude of indicators available to apply in evaluations, it is still the "famous" ones like the h-index, h, or citations per paper, CPP, that are predominantly used. Given the development in indicator construction, lesser-known indicators have been proposed to offer contextually appropriate solutions, confirming further that the famous indicators are not the best ones to use across the board in author-level evaluations, i.a. (Harzing & Alakangas, 2016; Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015; Wildgaard, 2015a; Wildgaard, 2015b). Therefore, this paper sets out to define an objective method that will help us sort through the wealth of available indicators, to identify a set of appropriate indicators that fit individual researchers with different academic profiles.

E-mail address: lorna.wildgaard@hum.ku.dk

http://dx.doi.org/10.1016/j.joi.2016.09.003
1751-1577/

Author-level indicators produce a single numerical value, which is used to describe a combined set of publication and citation data. This value typically depicts a central position within that set of data. Indicators are then a summary statistic, which calculate different measures of central tendency using a selected set of data within a dataset or all the available data. Parameters can be defined to determine how certain characteristics relevant to the data and researcher profile are calculated, for example weighting the age of citations or adjusting for citation velocity. Logically this means that under different conditions, because data has different characteristics, some indicators are more appropriate than other indicators. We need to learn more about under what conditions certain indicators are most appropriate. This is important work. The resulting indicator values are interpreted, in evaluations, as measures of the individual's prestige, production, or quality, etc., and they are used to inform decisions about an individual's academic future and as predictors of future performance. Several core developments have heightened the need for intensive action from the bibliometric community, to recommend indicators that support progressive evaluation and support the objectives of the researcher being evaluated:

1. Bibliometrics have become central to economic, political, social and academic evaluation systems, as well as to the individual profile of the researcher (Bloch & Schneider, 2016;; Sivertsen, 2016). Performance and assessment culture has been internalized and institutionalized in university and research institutions and consequently author-level indicators are being used as (self) regulatory tools to monitor and adjust scientific activities, in attempts to optimize the effect of the researcher and their publications (Wouters, 2014; Retzer & Jurasinski, 2009). Because of the critical issues of efficiency and trust in the evaluation system, (Abramo, D'Angelo, & Rosati, 2014), experts on bibliometric indicators do not generally see author-level indices as indicators of research quality. However, socially they seem to partly function like it (van Arensbergen, 2014).
2. Given the diversity of publication and citation cultures within scientific disciplines, the usefulness of an indicator is fluid (Lancho-Barrantes, 2010). An indicator that works well for one particular community of researchers is not necessarily appropriate in another community (apples to oranges) unless well-argued scaling factors are applied when measuring and comparing (Díaz-Faes, Costas, Galindo, & Bordons, 2015); (Abramo, Cicero, & D'Angelo, 2013). As of yet, there is no unified agreement on what these scaling or normalization factors should be and the invalidity of normalization continues to be discussed (Glänzel & Moed, 2013). Thus, bibliometric evaluation of the individual remains framed by culturally influenced norms, disciplinary norms, and "ways of knowing" in the individual's specialty, which is also affected by the individual's visibility or coverage in generic citation databases Harzing and Alakangas, 2016.
3. A prerequisite of informed bibliometric evaluation is that assessors understand the mathematical construction of the indicator and understand how well the mathematical model fits the data used to compute the indicator (Glänzel, 2010). This in turn improves understanding of how the indicator on a particular individual's publication/citation dataset serves as an asset or drawback in summarizing numerically the experiences and achievements of the researcher (Abramo & D'Angelo, 2011; Sandström & Sandström, 2009). The bibliometric community has for many years warned about the volatility of bibliometric statistics at the individual level. Of particular concern is the stability of the indicators and the importance of the numbers they produce, as they are based on limited data. The indicators are only informative with great methodological care (Hicks et al., 2015; IEEE, 2013; Bach, 2011).
4. At the individual level, disciplinary and personal culture has implications for the robustness and appropriateness of the bibliometric indicator in evaluations (Glänzel & Moed, 2013). Indicator values are influenced across disciplines and within academic ranks by the age, nationality, specialty of the researcher, length of career, amount of publications, publication language, number of collaborators, and position on the author by-line (Díaz-Faes et al., 2015; Levitt & Thelwall, 2014;; Claro & Costa, 2011;; Costas, van Leeuwen, & Bordons, 2010; Vinkler, 2007; Archambault & Gagné, 2004). Not to mention the availability of publication and citation data (Meho & Yang, 2007) and method of data-collection (Retzer & Jurasinski, 2009). The aforementioned are vital, non-consistent variables that differ from discipline to discipline, researcher to researcher and their influence must not be under-estimated in a useful and insightful bibliometric evaluation.

Notwithstanding the above complexities of individual bibliometric evaluation, author-level indicators are increasing in popularity, both in invention and in use by bibliometricians, researchers and administrators. Yet it falls to the responsibility of the bibliometric community to identify the stable indicators from the volatile and the true indicators from the spurious. The methodology used to recommend indicators has to be transparent and reproducible, ensuring that the principles for recommending appropriate indicators are not cloaked in math or reliant on fuzzy data. Thus, the research questions are these:

Does cluster analysis provide a useful method to identify disciplinary appropriate author-level indicators?

Does cluster analysis provide a useful method to identify seniority appropriate author-level indicators?

This paper investigates if the applied cluster methodology actually provides an informative approach in grouping researchers based on author-level indicator values. Can we draw informed observations or are the results purely arbitrary? Cluster analysis is a process-based methodology that to give meaningful results builds on seven stages, Bacher et al., 2010. Fig. 1 illustrates these stages which include: 1) data-collection, 2) description of data, 3) presentation, calculation and statistical description of bibliometric indicators, 4) a rationalized choice and application of the cluster algorithm and clustering statistics, 5) presentation and statistical description of clusters, 6) tests of the stability and strength of the clusters and finally, 7) informed interpretation of the clusters. These stages are used to organize the paper as follows: Stages 1 and 2 introduce the Methodology section, followed by a brief presentation of the bibliometric indicators, including their calcu-

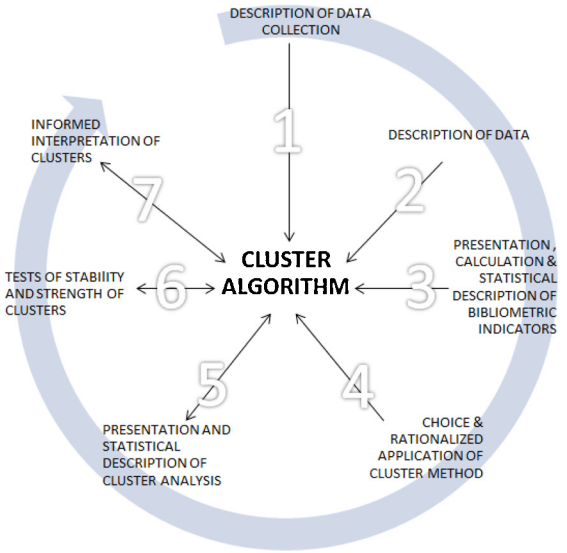

Fig. 1. Flow of the seven stages of a process-based cluster analysis tailored to bibliometrics.

lation and a statistical description, Stage 3. As it is important to be acquainted with structure of the data before choosing the clustering method, the background discussion in the Choice of Clustering Method section comes relatively late in the paper, in Stage 4. In this stage, cluster methodologies in statistical papers and in bibliometric studies are reviewed with the purpose to further rationalize the choice of cluster methodology. The Results section includes stages 5 and 6. Finally, Stage 7 is the Discussion, and finally conclusions drawn, suggesting directions for future research on this topic.

2. Data collection and methodology

2.1. Stages 1 & 2: data collection and description of data

In a previous study, researchers in Astronomy, Environmental Science, Philosophy and Public Health, were invited to take part in a survey about their academic profiles and web-presence, described in (Wildgaard, Larsen, & Schneider, 2013). From the information provided in the survey, it was possible to collect current curriculum vitae (CV) and publication lists (PL) for 741 of these researchers, (Wildgaard & Larsen, 2014). Publication and citation data was retrieved from Web of Science (WoS). The Centre for Science and Technology Studies (CWTS) at Leiden University, kindly supplied additional publication and citation information on articles and reviews in this dataset, from their custom version of the WoS. As neither the CWTS database nor the WoS database contained comprehensive data on Conference Proceedings, it was not possible to identify additional data on 3693 citable conference papers, thus these were excluded from the present analysis. In total 12,359 publications and 321,443 citations were retrieved in Astronomy; 7820 publications and 118,573 citations in Environmental Science; 3494 publications and 19,279 citations in Philosophy; and 7294 publications and 114,794 citations in Public Health, a summary is presented in Table 1. The sample consists of researchers employed in 24 European countries, (50% coming from the UK, Italy and Germany), 580 male researchers (78%) and 161 females (22%). The gender ratio reflects the SHE 2015 figures for gender in research and innovation, where it is reported that female researchers make up between a fifth and a third of the academic workforce, dependent on the research area (Antelo, 2016). Disciplinary and seniority information, provided by respondents to the survey, was used to group the researchers.

This unique dataset is composed of CVs and publication and citation data from four very different disciplines, each with very different sociological traditions regarding the activities of publishing and citing. The data represents junior and senior researchers, star researchers and mediocre alike. Importantly all with very different publishing histories. This composition is used to compare indicator values and cluster membership to each researcher's CV, allowing us to learn more about the rationality of implementing cluster methodology as a means to identify appropriate disciplinary and seniority indicators.

2.2. Coverage and publication characteristics at the disciplinary level

Comparing the CVs and publication lists (PL) and to the publications found in WoS revealed the following disciplinary trends:

1. *Astronomy* has a strong preference for multi-authorship on primarily article and conference papers, 12,359 out of 21,109 total publications reported on the researcher CVs/PL were identified in WoS, (58%).

Table 1
Distribution of publications identified in WoS. Citations identified in WoS and CWTS for 741 researchers across disciplines and seniorities.

		Publications		Citations	
Discipline	Sample	Median	IQR	Median	IQR
Astronomy, 192 researchers					
PhD	15	7.0	10.0	150.0	194.0
Post Doc	48	19.5	23.0	201.5	523.7
Assis Prof	26	39.5	40.7	702.0	1238.5
Assoc Prof	66	61.5	68.0	1214.0	2336.0
Professor	37	90	88.0	1889.0	2748.5
Environmental Science, 195 researchers					
PhD,	3	4.0	3;5[a]	34.0	16;60[a]
Post Doc	17	9.0	9.0	41.0	36.00
Assis Prof	39	18.0	12.0	148.0	271.0
Assoc Prof	85	29.0	35.0	326.0	600.5
Professor	51	64.0	57.0	692.0	1467.0
Philosophy, 222 researchers					
PhD	8	1.0	2.5	0.5	8.5
Post Doc	22	4.0	6.5	8.0	10.2
Assis Prof	43	6.5	9.7	6.5	27.2
Assoc Prof	74	7.0	8.0	8.0	36.5
Professor	75	18.0	26.0	29.0	92.0
Public Health, 132 researchers					
PhD	9	8.0	10.0	60.0	76.0
Post Doc	14	11.0	6.25	80.5	185.7
Assis Prof	30	22.0	24.0	167.2	354.2
Assoc Prof	50	43.0	53.7	518.0	877.5
Professor	29	76.0	93.0	954.0	2437.5

[a] minimum and maximum values.

2. *Environmental Science* publishes a great amount of article, conference papers and EU project reports; 7820 publications were identified in WoS, out of a total 16,720 publications listed on CVs/PL (47%).
3. *Philosophy* is a dialogue-based discipline, preferring single authorship publishing in blogs, in the media, books and in national languages, 3494 publications were identified in WoS, out of 14,724 publications listed on CVs/PL, (24%).
4. *Public Health* has a strong tradition of publishing articles in international journals in collaboration with medical researchers. They publish many articles and reports in local journals in national languages on local health issues and regulations; 7294 out of 9067 publications were identified in WoS, (80%).

2.3. Stage 3a: presentation of bibliometric indicators

The 44 indicators tested in this investigation are designed to measure different aspects of a researcher's productivity, visibility, currency, impact, prestige and collaboration. A full description of the 44 bibliometric indicators included in this analysis is presented in Appendix A. A summary is presented below.

1) *Publication-based indicators* indicate the productivity of the researcher *P* and *Fp*. While *App, mean pp collab* and *mean pp int collab* indicate the extent of collaboration by extracting information from the author bylines of the analyzed articles.

2) Citation-based *indicators:*
   (a.) *Citation count:* indicate the visibility or effect of the researcher's publications within their academic specialty. Effect is counted as citations, as in *C*, *CPP*, *Csc, sc, nnc, SIG,* and *Cless5*. *AWCR, Cage* and *PI* are adjusted for the age of the publications, while *Fc, FracCPP* and *AWCRpa* normalize for the number of authors written on the author byline of each paper.
   (b.) *Citation count normalized to publications and field:* indicators that compare the researcher's citation count to expected performance in their chosen field, *sum pp top n cits, sum pp top prop*, *NprodP* and *T > ca.*
   (c.) *Effect of output as citations normalized to publications and portfolio:* indicators that normalize citations to the researcher's portfolio, *%sc* and *%nc*.

3) *Indicators that indicate prestige using Journal Impact measures*: impact indicators of the journals a researcher has published in. These are used to suggest the researcher is cited more than average, and is of a high international standard, *mcs, mncs, mean mjs mcs, max mjs mcs and mean mjs*. Journal categories in the citation index, are used as a proxy for scientific fields.

Table 2
Comparison of median indicator scores, each cell shows median difference between disciplinary indicator values, minimum and maximum indicator value.

	Astronomy Med(min;max)	Environmental Science Med(min;max)	Philosophy Med(min;max)	Public Health Med(min;max)
Astronomy	–	1.7(0.5;5)	7.1(0.1;118)	1.6(0.2;5)
Environmental Science	0.6(0.2;1.7)	–	4.3(0.2;33)	0.9(0.5;1.5)
Philosophy	0.1(0;6.2)	0.2(0;3.5)	–	0.2(0;1.8)
Public Health	0.7(0.2;3.3)	1.1(0.6;2)	5.0(0.2;31.5)	–

4) *Hybrid indicators:* indicators of the level and performance of all of the researcher's publications or selected top performing publications. These indicators rank publications by the amount of citations each publication has received and establish a mathematical cut-off point for what is included or excluded in the ranking. They can be subdivided into indicators that are:
   (a.) *dependent on the calculation of the h index*: *h*, *ħ*, *Q2*, *h2*, *m*, *A*, the *e*-index (which supplements the *h*-index by computing the value of highly cited papers) and *hg* which allows a greater granularity in comparison between researchers with similar *h*- and *g*- indicators.
   (b.) *h "independent" indicators*: *AW* is the age-weighted indicator suggested for comparison with the *h*-index and the *g*-index allows greater distinction between the order researchers.
   (c.) *h adjusted to field: hnorm* allows across field comparison for multidisciplinary researchers,
   (d.) *h adjusted for co-authorship*: POPh, and,
   (e.) *h-type indicators of impact over time:* indicators of the extent a researcher's output continues to be used or the decline in use, *AR index. m*-quotient and *mg*-quotient are respectively the *h* and *g* indices divided by the academic age of the researcher.

2.4. Stage 3b: calculation of bibliometric indicators

Thirty-seven out of 44 bibliometric indices were calculated manually using Microsoft Excel. The field-level performance indicators, *sum pp top n cits, sum pp top prop, mcs, mncs, mean mjs mcs, max mjs mcs* and *mean mjs* were adjusted to the individual-level and supplied by CWTS. For both calculation methods each individual researcher's number of publications and citations found in WoS were used. "Academic age" is used in a number of indicators as a normalization factor. It is calculated as the number of years since the researcher's first publication registered in the WoS subtracted from 2013 (the year of data-collection).

2.5. Stage 3c: statistical description of bibliometric indicators

IBM SPSS version 22 was used for the cluster analysis and calculation of statistics. Boxplots for each indicator can be found in Appendix B, provided as Supplementary material. Indicator values are not normally distributed within the disciplines, typically right-skewed with an overweight of lower indicator values. Astronomy scored on average (median) a ratio of 1.7 and 1.6 times higher than Environmental Science and Public Health respectively, and on average 7.1 times higher than Philosophy, Table 2.

In summary, Table 3 provides an overview of median, minimum and maximum indicator values and the interquartile range. Astronomy researchers had the highest median indicator values on 37 out of the 44 indicators, followed by Public Health, Environmental Science and Philosophy, scoring the lowest indicator values. Astronomy scored the highest median values on all of the Publication-based indicators and all but one of the hybrid indicators. This indicator was the *NprodP* where Philosophy had the highest median number of productive papers, followed, in descending order, by Public Health, Astronomy and Environmental Science. In regards to the Citation-based indicators Philosophy had the highest median values for *%nc*, followed by Environmental Science, Public Health and Astronomy; Public Health had the highest median number of non-cited publications (*nnc)* with Philosophy, Environmental Science and Astronomy scoring the same median value. The *Cage* indicator showed publications from Environmental Science being cited on average the quickest, followed by Astronomy, Public Health and Philosophy. *PI* showed only a marginal difference between Public Health and Astronomy, Philosophy and Environmental Science returning the same median values. Concerning the group of indicators that measure "Prestige using journal impact measures", Astronomy outperformed on average on the indicators *Mcs, Mean mjs mcs* and *max mjs mcs.* The indicators *mncs* and *mean mnjs* returned the same median values across all disciplines.

Worth noting is even when median indicator values were very similar, or the same, across all four disciplines, the variability of the scores within each discipline were vastly different. This variability reflects characteristics of the individual researcher, which are important to capture in the cluster analysis in order to create meaningful groups of researchers.

Table 3
Basic description of indicator values at the disciplinary level. Reported values: median, interquartile range (IQR), minimum (min) and maximum (max).

Type.	Indicator	Astronomy				Environmental Science				Philosophy				Public Health			
Publication		median	IQR	min	max	median	IQR	min	max	median	IQR	min	max	median	IQR	min	max
	P	46	61	2	327	26	45	1	425	9	14	1	140	30	53	1	661
	Fp	15	25	.5	137	11	18	.5	137	8	12	.5	140	11	19	.5	155
	App	5	4	1	10	3	1	1	7	1	.3	1	5	4	2	1	8
	Mean pp collab	1	.2	0	1	.5	.3	0	1	0	.3	0	1	1	.2	0	1
	Mean pp int collab	1	.2	0	1	.2	.3	0	1	0	.1	0	1	.2	.3	0	1
Citation	*C*	740	1881	3	16481	273	611	0	14141	15	30	0	100	313	815	0	13520
	Csc	499	1170	1	12605	183	454	0	11750	9	34	0	2934	249	624	0	11030
	Sc	236	592	1	6188	67	160	0	2391	2	9	0	809	63	197	0	2490
	%sc	34	16	10	77	23	17	0	87	15	30	0	100	20	15	0	67
	Nnc	4	7	0	36	4	5	0	33	4	6	0	114	8	11	0	173
	%nc	8	9	0	67	14	15	0	100	50	44	0	100	27	21	0	100
	CPP	17	16	1	59	9	10	0	51	1	3	0	34	10	10	0	74
	Sig	106	153	2	1365	42	79	0	1378	6	14	0	360	61	111	0	1040
	Cless5	512	1142	0	10704	188	371	0	6958	8	28	0	1575	235	499	0	8230
	PI	71	21	0	100	66	27	0	123	66	41	0	100	72	30	0	100
	Cage	3	2	0	7	3	2	0	8	1	2	0	13	2	1	0	6
	Fc	248	577	1	6734	100	287	0	4494	11	37	0	1895	102	244	0	2797
	FracCPP	13	12	0	64	7	8	0	29	0.3	4	0	59	8	7	0	63
	Sum pp top n cits	16	30	0	192	5	15	0	235	0	1	0	65	6	17	0	243
Hybrid	*H*	15	16	1	66	9	9	0	59	2	3	0	32	9	11	0	60
	AWCR	106	219	0.7	2219	38	72	0	1425	2	6	0	354	55	90	0	1882
	AWCRpa	38	63	0.3	498	14	31	0	400	2	6	0	163	18	39	0	423
	G	26	28	1	119	14	17	0	99	3	5	0	56	17	18	0	93
	H2	9	7	1	25	6	5	0	24	2	2	0	15	7	5	0	23
	POPh	6	6	0	33	4	5	0	27	1	2	0	17	4	5	0	19
	M-quot	1	1	0.1	5	0.6	0.4	0	2	0.2	0.2	0	2	0.8	0.8	0	3
	Mg-quot	2	2	0.2	7	1	1	0	3	0.2	0.4	0	2	1	1	0	5
	Q2	21	22	1	96	12	14	0	82	3	4	0	43	14	13	0	78
	AW	10	10	0.8	47	6	6	0	38	1	2	0	19	7	6	0	43
	E	18	19	1	79	10	11	0	57	2	4	0	39	13	13	0	67
	M	29	29	2	150	18	17	0	115	4	7	0	94	20	20	0	121
	A	38	41	2	210	21	25	0	150	5	8	0	94	27	32	0	199
	AR	6	3	1	14	4	2	0	12	2	2	0	10	5	3	0	14
	Hg	19	22	1	89	11	13	0	76	2	4	0	42	12	15	0	74
	hnorm	0.5	0.4	0	7	0.4	0.3	0	4	0.3	0.3	0	2	0.4	0.4	0	3
	ħ	19	22	1	90	12	12	0	84	2	4	0	41	12	14	0	82
	Sum pp top prop	3	9	0	80	2	5	0	115	0	1	0	30	2	6	0	100
	NprodP	22	45	0.2	260	20	44	0	580	24	83	0	2718	22	53	0	834
	T>ca	2	2	0.2	22	1	1	0	6	0.2	0.6	0	4	1	1	0	6
Prestige	*Mcs*	12	11	0.3	58	7	10	0	50	2	3	0	53	10	12	0	98
	mncs	1	0.8	0	5	1	0.6	0	4	1	1	0	25	1	0.8	0	4
	Mean mjs mcs	14	9	0.7	37	9	8	0.2	37	2	4	0	78	10	10	0	28
	Max mjs mcs	42	78	1	320	30	26	0.4	289	5	13	0	277	34	44	0	261
	Mean mnjs	1	.3	.2	2	1	.5	.1	2	1	.7	0	8	1	.3	0	2

3. Choice of clustering method

3.1. Stage 4: choice and rationalized application of the cluster algorithm

Clustering and mapping techniques have similar objectives and terminology and are often used together in bibliometric analyses, in visualization of collaborations and research areas, in the development of new bibliometric software and in exploring the overlap and redundancy between indicators. However, these techniques are based on different ideas and rely on different assumptions (Waltman, Van Eck, & Noyons, 2010). This paper concentrates solely on cluster analysis.

Cluster analysis is concerned with exploring data and finding structure in this collection of elements that are characterized by a number of variables. The aim is to group the elements in this collection so that they are grouped in "Clusters". Each cluster contains very similar elements and is preferably highly heterogeneous from the elements grouped in the other clusters. From this clustering, the internal structure of a dataset according to some chosen attributes can be interpreted. It is a tool for researchers to understand groupings of data and gain knowledge of how to classify elements in multidimensional datasets by interpreting their similarities and dissimilarities. The four main frameworks for cluster analysis are *probabilistic, partitioning, hybrid* and *hierarchical* clustering which are described in technical detail for example in (Ibáñez, Larrañaga, & Bielza, 2013); (Bacher, Pöge, & Wenzig, 2010). Within each framework there have been developed hundreds of different clustering algorithms by researchers from different scientific disciplines (Äyrämö & Kärkkäinen, 2006). Each framework uses a different starting point for creating clusters. Consequently, as each clustering algorithm is different, different ordination and

clustering results are produced when used on the same data. Thus clustering results can be considered arbitrary (Schneider & Borlund, 2007a;; Schneider & Borlund, 2007b). Therefore, it is important to justify the chosen clustering method, as the method can influence the validity of the results and the conclusions drawn from these. In this paper, a two-step clustering method is used. This choice is rationalized and motivated in the following through: 1) reference to clustering techniques implemented in previous bibliometric studies and, 2) to technical issues of clustering as argued in statistics literature.

The dataset used in this paper consists of small sets of bibliometric data that contain skewed, mixed scale data, that is nominal, ordinal and interval data together. It is thus a requirement that the chosen clustering method can cluster skewed data on multiple scales, for both researchers (741 cases) and indicators (44 variables). Therefore, the *Probabilistic Clustering* method, the *Partitioning Clustering* method and the *Hybrid Clustering* method are not appropriate methodologies. The reasons for their elimination are described briefly in the following.

The *Probabilistic* approach allows for uncertainty in cluster membership to enable identification of possible sub-components of, for example, collaboration, and can be used to provide practical insights for policymakers by creating a taxonomy of collaboration and characteristics of type (Jeong & Choi, 2012). Even though the *Probabilistic* method supports clustering of data on different scales (without the transformation of variables), it only allows the clustering of cases. The size of the dataset needs to be large, e.g. n = 3000, so as to assign a case the probability of belonging to a cluster based on patterns in the data (Bacher, 2000). The *Partitioning* method is simple and its computational efficiency make it a common approach in bibliometric studies, i.a. identifying disciplinary research priorities and cooperation (Chawla, 2006) or bibliographic links between journals, distance metrics and impact factors (Su, Shang, & Shen, 2013). Although useful on moderately sized sample sizes (e.g. N = 300), it again only clusters cases, and not variables. Finally, the *Hybrid Clustering* method associates each data element with a set of membership levels that indicate the strength of the association between that element and a particular cluster. As a result, the element can belong to more than one cluster, and outliers and elements that link clusters together are identified (Janssens, Zhang, & Glänzel, 2009). This method has proved useful in producing less arbitrary clusters than methods that attempt exclusive clustering (Ruspini, 1970). It has been used in classification schemes such as the Essential Science Classification that forms part of the Web of Science, where cluster analysis and cognitive mapping are integrated into subject classification (Janssens et al., 2009). Yet the *Hybrid* method is not chosen for this paper because finding the optimal cluster number and identifying why an element is placed in a particular cluster, can give ambiguous results.

We are striving for exclusivity in the clusters, using data does not fulfill the assumption of normality. The dats is mixed-scaled and the sample size is small. Therefore a *Hierarchical Clustering* approach is considered appropriate Bacher et al., 2010. Such methods: have been used by (Otsuki & Kawamura, 2013) who combine co-citation and bibliographic coupling with regional data and purchase information from Twitter to visualize purchasing behavior; and, similarly (Sun et al., 2014) who identify core target journals to create an overview of a discipline's key domains, indicating areas for further research and development. Working with data on different scales present difficulties. Traditional hierarchical clustering requires the data downgraded to nominal measures and matching type coefficients used to form clusters in the structure of the data (Gower, 1967). However, the *Two-Step* hierarchical clustering algorithm allows the cluster of cases and variables, recommending but not requiring normalization, and it enables the analysis of a small sample of mixed scale data. Further, the *Two-Step* cluster eliminates strong links with other clusters that distort the intra-cluster coherence, so it produces more robust clusters than the *Hybrid* method and unlike the *Partitioning* and *Probabilistic* algorithms, *Two-Step Clustering* produces a hierarchical structure of clusters, enabling the identification of parental relationships between bibliometric indicators on a relatively small amount of data (MacQueen, 1967). It has the flexibility to merge smaller clusters into larger ones (agglomerative clustering) or split larger clusters (divisive clustering) dependent on the character of the data, as exemplified in (Ibáñez et al., 2013).

Based on the aforementioned discussion of the advantages and disadvantages of clustering techniques, the *Two-Step Clustering* method is chosen. It is arguable the appropriate approach for the type of data in our dataset and conforms with the aim of the analysis. We choose not to normalize the data, aiming to preserve transparency and methodological simplicity.

4. Results

IBM SPSS version 22 was used for the cluster analysis and calculation of statistics. A statistical description of each indicator was performed. This was done to explore the range and spread of indicator scores and consequently aid interpretation of how indicator scores summarize the performance of the researchers. A Two-Step cluster analysis followed to segment the researchers. The data was not transformed as the clustering algorithm is thought to behave reasonably well when the assumption of normality is not met and we are striving for a simple, transparent model (Kaufman & Rousseeuw, 2005). We are aware that we could have chosen to transform the data by computing the inverse, the square root or logarithm etc., in an attempt to find an appropriate normalization and supplement this with a good mathematical argument for treating all the data in the same way. However, transformation can conceal valuable data and it is typical for bibliometric data that the interesting cases lie in the tails of the distribution. The non-parametric nature of the data is a characteristic we choose not to manipulate. Likewise, we did not adjust for outliers in the definition of the cluster model. The distance measure log-likelihood = 1 was used, as this is suitable for both continuous and categorical variables. The Bayesian Information Criterion (BIC) was chosen to describe the relations in the data. BIC performs well when maximizing discrimination in smaller datasets, producing a simple model with the least assumptions and variables but with greatest explanatory power (IBM, 2012). As the final cluster solution may depend on the order of the cases in the file, the cases were arranged in random order. Although the

Table 4
Segmentation of researchers, 4 cluster solution.

		ASTRONOMY		ENVIRONMENTAL SCI.		PHILOSOPHY		PUBLIC HEALTH	
Cluster	Description	Size	Academic Age	Size	Academic Age	Size	Academic Age	Size	Academic Age
1	Below interquartile range	53	7 (2;18)	63	9(2;31)	132[a]	9 (1;33)	59	9 (1;25)
2	Median interquartile range	58	21 (10:33)	86	17 (7;36)	72	15 (3;33)	48	15 (4;34)
3	Top of interquartile range	62	15 (3–34)	45	22 (9;34)	14	18 (7–33)	21	21 (10;33)
4	Extreme outliers	19	23 (11–33)	1	32	4	22 (15;30)	4	22 (19;28)

[a] low values were below median but within interquartile range.

lowest BIC coefficient was for five clusters, according to the SPSS algorithm, the optimal number of clusters is four, because the largest ratio of distances was for four clusters. The resulting four-cluster solution divided the researchers into four groups: extremely high, high, middle, low scores on the bibliometric indicators. With reference to the model summary statistics in the SPSS model viewer, the 4 cluster-solution was deemed fair to good across all disciplines and cluster distribution. *F*-test statistic was used to state the statistic importance of each indicator as a predictor of a researcher being placed in a specific cluster. The limitations of significance tests in research assessment have been well documented (Schneider, 2013), therefore the practical importance of the difference between clusters, was assessed using Hedges' *g* (Hedges, 1981). Hedges' *g* provides a superior estimate of the relative magnitude of the difference between means in small clusters of dissimilar size, by weighting each cluster's standard deviation by its sample size, thus correcting the bias of the more commonly used Cohen's *d* (Cohen, 1988). Effect sizes were computed using the effect size software developed by (Ellis, 2009). Then the mean value of each indicator was used to summarize similarities and dissimilarities between clusters within each discipline. Odds ratios were calculated to analyze the likelihood of a researcher of a specific academic seniority being placed in a specific cluster and these results also informed deliberations of the extent coverage or indexing practices in the WoS distorted interpretation of researcher prestige. Finally, correlations of researchers measured on complementary indicators were performed. This was done to visualize the difference between and within clusters and to study changes in researcher rankings.

4.1. Stage 5: presentation and statistical description of clusters

Table 4 shows all researchers grouped in the four clusters, the number of researchers in each cluster (size) and the mean academic age of the researchers in each group. The minimum and maximum academic ages are shown in parenthesis.

The cluster comparison function in the SPSS model viewer, illustrates the distribution of values within each cluster. The distributions are overlaid on a boxplot for the distribution of values overall. This showed that the indicator scores for researchers grouped in:

Cluster 1, were low scores below the interquartile range on the left tail of the distribution.

Cluster 2, were scores evenly distributed between the median and the upper line of the boxplot.

Cluster 3, were scores that lay a little higher than the interquartile range, and approached a normal distribution, while

Cluster 4, were the extremely high scores in the right tail of the distribution.

Table 5 shows the mean indicator scores within the four clusters for each discipline. The majority of indicator scores increased across the clusters. It was the same indicators, which increased in value across all disciplines. The indicator values for percent self-citations *(%sc)* and percent not cited *(%nnc)* decreased across clusters, that is the low scoring researchers in Cluster 1 had the highest proportion of un-cited papers and self-citations. This was expected as the composition publications and citations used to calculate these ratio indicators increased from Cluster 1 to Cluster 4. Across all disciplines, PhD students and Post Doc researchers were dominant in Cluster 1, and seniority increased in Cluster 2 and 3 along with the academic age of the researcher. Consequently, *hnorm* generally decreased across clusters, as this indicator calculates the proportion of productive papers, *h,* to the total number of papers a researcher has produced. As the researchers in Cluster 1 had fewer publications, the ratio productive papers to total papers was smaller and vice versa for researchers in Cluster 4. However, because *hnorm* normalizes *h* in this way, it enabled comparison across disciplines, (Levitt and Thelwall, 2014), and showed that the Philosophers in Clusters 1, 2, 3,and 4 had on average a very similar number of citations per paper as their cluster counterparts in Astronomy, Environmental Science and Public Health. The extreme indicator values in Cluster 4 were attributed to Associate Professors and Professors who, across all disciplines, ranked the highest scores in *C, sc, AW, h, ħ, Cless5, Csc, sum pp top n cits, hg, Q2, h2 and Nprod.* Meaning that these researchers, although they did not necessarily publish the most, were cited the most, were cited quickly and had the most impactful papers even when adjusted to the performance of their entire portfolio.

The researchers in Cluster 3, the high scorers, come second to Cluster 4 researchers when ranked according to the hybrid *indicators,* (*h, g, h2, Q2, hg,* etc.), but scored higher on field-normalized indicators such as *mncs, sum pp top prop* and *mean mjs mcs, max mjs mcs.* Further, they achieved a higher rank placement than Cluster 4 researchers when the *SIG* indicator was used i.e. the most significant paper, showing that they had one very high scoring publication, higher than citations to papers produced by Cluster 4. Cluster 2, the middle performers, scored well on indicators that normalized or rewarded authorship, *FracCPP, mean pp collab* and *mean pp int collab.* This cluster ranked generally lower than Cluster 3 across the hybrid indicators but using the *m index,* i.e. the median number of citations per paper in the *h*-index, increased their rank position above Cluster

Table 5
Mean indicator scores within the four Clusters (1 = low, 2 = median, 3 = top, 4 = extreme) per discipline. Indicators used in projection figures are marked in bold.

	Astronomy				Environmental Sci.				Philosophy				Pub. Health			
Cluster	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
P	15.9	53.4	70.4	213.2	10.6	37.6	77.3	425.0	6.4	24.1	35.9	99.7	17.3	48.2	140.7	249.7
Fp	**5.9**	**26.8**	**24.6**	**60.1**	**5.2**	**16.5**	**34.7**	**136.2**	**6.2**	**22.1**	**25.2**	**61.9**	**7.6**	**17.1**	**41.9**	**68.9**
App	4.1	3.6	5.4	6.6	2.9	3.3	3.5	3.9	1.1	1.5	2.2	2.6	3.5	4.4	5.0	4.6
Mean pp collab	0.7	0.7	0.8	0.8	0.5	0.5	0.5	0.4	0.1	0.2	0.5	0.5	0.6	0.6	0.7	0.6
Mean pp int collab	0.7	0.7	0.7	0.8	0.2	0.3	0.3	0.3	0.0	0.1	0.2	0.4	0.3	0.3	0.4	0.3
C	**150.9**	**741.7**	**1990.7**	**7736**	**46.6**	**361.4**	**1564.6**	**14141.0**	**5.6**	**65.5**	**423.7**	**1971.2**	**97.5**	**591.3**	**2602.9**	**6499.2**
Csc	88.7	484.9	1357.6	5133.1	33.8	259.9	1237.5	11750.0	4.1	53.1	331.3	1528.2	72.1	468.9	2077.7	5593.2
Sc	62.2	256.8	633.1	2.603	12.7	101.5	327.0	2391.0	1.4	12.4	92.3	443.0	25.3	122.3	525.1	906.0
%sc	40.6	34.8	33	35.7	27.0	27.7	21.0	16.9	17.8	20.4	22.1	22.9	25.4	19.4	20.0	11.7
Nnc	2.1	5.6	4.7	14.3	2.6	5.6	7.6	33.0	4.1	12.8	8.7	18.2	6.0	12.1	30.9	69.7
%nnc	14.7	10.9	6	6.2	24.9	14.8	9.4	7.7	64.7	41.4	24.9	18.1	37.7	24.6	20.6	29.7
CPP	**9.0**	**14.0**	**30.2**	**36.6**	**4.7**	**10.2**	**21.6**	**33.2**	**0.8**	**3.5**	**14.2**	**19.7**	**5.2**	**14.1**	**20.1**	**45.6**
Sig	34.2	98.8	231.0	566.0	15.7	57.9	198.6	503.0	2.9	19.7	140.0	220.7	26.8	105.3	269.8	810.2
Cless5	122.0	453.9	1325.1	5013.8	304.7	316.6	439.5	21.0	3.8	38.2	230.5	1044.5	74.6	394.8	1500.2	3743.5
PI	80.9	62.3	69.5	66.6	79.1	65.6	55.2	49.2	55.0	65.5	57.6	54.1	79.9	67.5	58.3	49.5
Cage	1.6	3.7	3.0	3.4	2.0	3.1	4.2	4.9	1.2	2.1	3.1	3.5	1.4	2.4	3.2	3.0
Fc	**49.9**	**326.7**	**664.7**	**2089.8**	**21.9**	**150.7**	**707.4**	**4494.5**	**5.3**	**54.3**	**211.0**	**1230.2**	**37.7**	**192.9**	**753.0**	**1514.8**
FracCPP	6.8	11.2	22.4	28.6	3.5	7.4	17.2	27.4	0.4	4.2	16.7	14.5	6.3	10.2	13.7	33.7
Sum pp top n cits	2.8	15.2	31.7	102.0	0.9	7.5	32.9	235.0	0.0	1.4	7.8	36.0	2.2	13.4	52.8	89.2
h	**6.3**	**14.7**	**22.9**	**44.8**	**3.6**	**10.1**	**20.8**	**59.0**	**1.1**	**4.1**	**9.4**	**23.5**	**4.9**	**12.3**	**26.6**	**32.2**
AWCR	35.7	86.6	286.0	993.1	9.1	50.28	162.4	1425.1	0.8	8.2	47.0	204.7	19.1	84.4	299.5	781.4
AWCRpa	11.7	35.9	78.1	212.9	4.2	21.5	66.9	400.8	0.8	7.1	33.1	111.7	8.5	30.1	90.3	211.8
g	10.2	23.5	40.3	77.2	5.4	16.3	35.4	99.0	1.1	6.5	18.8	40.5	8.0	21.7	45.2	68.7
h2	4.8	8.6	12.1	19.3	3.3	6.8	11.3	24.1	1.3	3.7	7.3	12.2	4.1	8.1	13.6	17.5
POPh	2.5	6.7	9.2	14.8	1.7	4.7	9.9	27.0	0.6	2.7	5.2	13.0	2.3	5.4	10.1	13.2
m-quot	**0.9**	**0.7**	**1.7**	**2.0**	**0.5**	**0.6**	**0.9**	**1.8**	**0.1**	**0.3**	**0.6**	**1.1**	**0.6**	**0.9**	**1.3**	**1.5**
mg-quot	1.4	1.2	3.0	3.5	0.7	1.0	1.7	3.0	0.1	0.5	1.3	1.8	1.0	1.7	2.3	3.3
Q2	9.2	19.3	32.2	60.4	5.2	13.7	28.0	82.3	1.6	6.0	15.1	31.9	7.3	17.8	35.4	51.6
AW	5.3	8.8	16.1	30.8	2.8	6.8	12.4	37.7	0.7	2.7	6.6	13.8	3.9	8.8	17.0	25.0
e	7.6	16.0	29.7	52.6	4.2	11.4	25.1	59.6	0.9	4.9	15.3	28.7	6.1	16.4	31.7	52.6
m	**14.0**	**25.7**	**46.1**	**81.9**	**8.3**	**19.1**	**38.2**	**115.0**	**2.5**	**9.3**	**29.0**	**43.5**	**11.2**	**27.2**	**47.5**	**96.1**
A	**16.4**	**33.0**	**64.0**	**117.6**	**9.4**	**24.2**	**53.1**	**149.7**	**2.6**	**10.9**	**41.2**	**59.4**	**13.3**	**37.0**	**66.0**	**160.4**
AR	3.9	5.6	7.9	10.6	2.9	4.8	7.2	12.2	1.3	3.2	6.2	7.6	3.5	6.0	8.1	12.5
hg	8.0	18.6	30.3	58.7	4.4	12.8	27.1	76.4	1.0	5.1	13.1	30.8	6.2	16.3	34.6	46.3
hnorm	**1.2**	**0.5**	**0.8**	**0.3**	**0.7**	**0.5**	**0.5**	**0.1**	**0.2**	**0.4**	**0.5**	**0.3**	**0.5**	**0.5**	**0.2**	**0.3**
ħ	7.9	18.4	30.2	60.7	4.4	12.9	27.9	84,0	1.3	5.3	14.2	30.6	6.2	16.6	35.6	53.4
Sum pp top prop	0.9	2.4	10.8	36.7	0.3	2.6	10.1	115.4	0.1	1.9	4.5	15.5	1.0	4.7	16.1	35.6
NprodP	8.8	44.1	36.1	129.8	14.1	47.45	72.2	580.6	91.4	220.1	60.3	82.5	17.5	45.7	123.4	260.7
T > ca	2.5	1.5	4.2	2.6	1.5	1.3	1.8	1.2	0.3	0.6	1.3	1.5	1.4	1.9	1.8	3.1
Mcs	5.9	9.6	21.8	24.8	4.1	8.4	18.8	30.2	1.0	3.8	15.7	19.8	5.2	15.7	20.5	62.4
mncs	0.7	0.6	1.6	1.6	0.8	0.9	1.4	2.6	0.7	1.3	1.9	1.6	0.8	1.5	1.4	2.9
Mean mjs mcs	8.9	15.1	17.3	19.7	5.3	9.7	16.4	19.5	1.6	4.7	16.9	17.0	6.1	14.0	17.4	25.5
Max mjs mcs	23.2	61.4	105.6	187.9	13.8	34.1	75.7	111.5	3.6	18.9	98.1	121.8	18.4	52.9	111.7	180.3
Mean mnjs	1.0	0.9	1.2	1.3	0.8	1.0	1.2	1.6	0.8	1.2	1.5	1.2	0.9	1.2	1.2	1.3

3, but below Cluster 4, researchers. The researchers with low scores, Cluster 1, had the fewest publications and citations, and ranked below the researchers in Clusters 2, 3 and 4 across hybrid indicators. Removing self-citations or calculating *CPP* further reduced their scores, however using the *mg* or *m-quotient,* which normalize the *h* and *g* indices for the academic age of the researcher, or indicators of the currency of papers, *PI and Cless5,* raised the scores on a par with the researchers in Cluster 2 and in some cases higher than Cluster 3.

4.2. Stage 6a: tests of the stability and strength of the clusters

4.2.1. Indicator score as a predictor of cluster membership

The statistic importance of each indicator as a predictor of cluster membership was investigated. This analysis used the *F*-test statistic, where scores range between 0 and 1, the closer to 1 the less likely the variation for a variable between clusters is due to chance and more likely due to some underlying difference (IBM, 2012). The post hoc test Tamhane T2 was applied to compute the *F*-test statistic because equal variance in the data is not assumed and sample sizes are unequal. The following indicators were identified with predictor importance = 1: *h2-index* determined cluster membership in Astronomy, *sum pp top prop* in Environmental Science, *Q2* in Philosophy and the *e*-index in Public Health. Table 6 presents the indicators, which mean indicator values statistically discriminated between Clusters 1, 2, 3 and 4. These are ranked in descending order of importance for researcher inclusion in the cluster (001 alpha level):

Table 6
Indicators with statistically different mean values*.

Discipline	Indicator
Astronomy	*h2, ħ, g, hg, e, AR, Q2, h, AW, Cless 5, A, sc, m, C, AWCR, sum pp top n cits*
Environmental Science	*sum pp top prop, C, Csc, AWCR, sum pp top n cits, g, ħ, Q2, hg, h2, h, AW, e, POPh, AR, sc, A, AWCRpa, P, Fc, FracCPP, CPP, Fp, mcs, mean mjs mcs, mg, max mjs mcs, SIG, cage, PR, %nc*
Philosophy	*Q2, g, ħ,hg, e, Fc, AW, AWCRpa, h, Cless5, h2, C, AWCR, Csc, AR, sum pp top prop*
Public Health	*e, A, g, h2, m, ħ, Q2, AR, hg, Sig, AW, h, Csc, poph, C, Fc, AWCRpa, Cless5, AWCR, sum pp top n cits, sc, P, sum pp top prop*

*Note: Philosophy and Public Health only showed statistic difference between Clusters 1, 2 and 3. As Environmental Science had only 1 researcher in Cluster 4, this cluster was not included in the analysis.

Table 7
Hedges' g by cluster, within discipline.

Discipline	Seniority	Gender	Academic Age	Collaboration	Publications	Citations	h2	Sum pp top prop	Q2	e	hg	CPP
Astronomy												
Cluster 1 vs Cluster 2	−2.3	0	**−2.3**	0.3	**−1.7**	**−1.7**	**−2.2**	−0.7	**−1.9**	**−2.0**	**−2.1**	**−0.9**
Cluster 1 vs Cluster 3	−1.4	0	**−1.3**	−0.6	**−1.8**	**−2.3**	**−3.4**	**−1.9**	**−3.0**	**−3.3**	**−3.0**	**−2.7**
Cluster 1 vs Cluster 4	−0.2	0.3	**−3.2**	**−1.2**	**−5.1**	**−3.9**	**−7.4**	**−3.6**	**−6.2**	**−5.9**	**−7.0**	**−3.8**
Cluster 2 vs Cluster 3	0.3	0	0.7	**−1.0**	−0.5	−0.5	**−1.5**	**−1.6**	**−1.5**	**−1.9**	**−1.5**	**−2.1**
Cluster 2 vs Cluster 4	−0.6	0.3	−0.2	**−1.9**	**−3.7**	**−3.6**	**−5.0**	**−3.5**	**−4.7**	**−4.6**	**−5.0**	**−3.2**
Cluster 3 vs Cluster 4	**−0.8**	0.3	**−0.9**	−0.5	**−2.9**	**−2.7**	**−2.7**	**−2.3**	**−2.5**	**−2.3**	**−2.7**	−0.6
Environmental Science												
Cluster 1 vs Cluster 2	**−0.8**	0.2	**−1.1**	−0.4	**−1.6**	**−2.0**	**−3.2**	**−1.3**	**−2.7**	**−2.5**	**−2.7**	**−1.5**
Cluster 1 vs Cluster 3	**−1.4**	0.5	**−1.7**	−0.6	**−2.7**	**−3.2**	**−6.1**	**−3.1**	**−5.6**	**−4.7**	**−5.6**	**−3.0**
Cluster 1 vs Cluster 4	–	–	–	–	–	–	–	–	–	–	–	–
Cluster 2 vs Cluster 3	−0.7	0.3	−0.7	−0.2	**−1.4**	**−2.6**	**−3.2**	**−2.2**	**−3.1**	**−2.9**	**−3.1**	**−2.0**
Cluster 2 vs Cluster 4	–	–	–	–	–	–	–	–	–	–	–	–
Cluster 3 vs Cluster 4	–	–	–	–	–	–	–	–	–	–	–	–
Philosophy												
Cluster 1 vs Cluster 2	−0.4	0.2	**−0.7**	−0.7	**−1.0**	**−1.8**	**−2.6**	**−1.2**	**−2.7**	**−2.2**	**−2.7**	**−2.8**
Cluster 1 vs Cluster 3	−0.3	0.2	**−1.1**	**−2.9**	**−3.8**	**−8.2**	**−6.5**	**−4.8**	**−9.9**	**−8.4**	**−8.7**	**−5.4**
Cluster 1 vs Cluster 4	**−0.8**	0	**−1.8**	**−4.0**	**−13.6**	**−12.4**	**−11.5**	**−9.7**	**−18.2**	**−18.8**	**−19.2**	**−13.1**
Cluster 2 vs Cluster 3	0.1	0	−0.2	**−0.8**	−0.4	**−4.3**	**−3.8**	**−1.0**	**−4.1**	**−3.8**	**−3.4**	**−3.0**
Cluster 2 vs Cluster 4	−0.3	−0.3	−0.7	**−1.2**	**−2.8**	**−8.6**	**−8.6**	**−4.4**	**−9.9**	**−9.1**	**−9.7**	**−6.6**
Cluster 3 vs Cluster 4	−0.5	−0.2	−0.5	−0.3	**−3.0**	**−3.0**	**−3.7**	**−2.0**	**−4.0**	**−2.5**	**−3.7**	−0.6
Public Health												
Cluster 1 vs Cluster 2	**−0.9**	0	**−1.0**	−0.6	**−1.3**	**−2.2**	**−2.8**	**−1.6**	**−2.7**	**−2.9**	**−2.5**	**−1.9**
Cluster 1 vs Cluster 3	**−1.3**	0.4	**−2.3**	**−1.1**	**−3.7**	**−5.8**	**−6.7**	**−4.7**	**−7.0**	**−7.1**	**−7.1**	**−3.8**
Cluster 1 vs Cluster 4	**−1.6**	0.4	**−2.6**	−0.7	**−3.5**	**−5.5**	**−7.4**	**−1.1**	**−7.9**	**−11.0**	**−6.3**	**−5.9**
Cluster 2 vs Cluster 3	−0.6	0.4	−0.7	−0.4	**−2.1**	**−3.8**	**−3.8**	**−2.6**	**−3.6**	**−3.6**	**−3.7**	**−1.0**
Cluster 2 vs Cluster 4	**−1.1**	0.4	**−0.8**	−0.1	**−2.5**	**−4.5**	**−4.9**	**−2.7**	**−2.2**	**−7.2**	**−3.9**	**−3.5**
Cluster 3 vs Cluster 4	−0.5	0	−0.1	0.3	**−1.9**	**−1.8**	**−1.6**	**−1.1**	**−1.7**	**−3.1**	**−1.1**	**−2.2**

Effect size: g < 0.20 trivial, g ≥ 0.20 small effect, g ≥ 0.50 medium effect, g ≥ 0.80 large effect (in bold) and g ≥ 1.30 very large effect (in bold). These benchmarks must not be used uncritically.

4.2.2. Results: of effect size analysis

Table 7 displays the effect sizes of indicators, between clusters, within discipline, using Hedges' *g*. The data provides a great choice of indicators and combinations between and within clusters. Therefore, the demographic variables (*seniority*, *gender* and *academic age*) and indicators that are of typical bibliometric interest in researcher comparisons are presented, (number of co-authors defined as "*collaboration*", number of publications, *P*, number of citations, *C*, and citations per paper, *CPP*). Likewise, the hybrid indicators that were predicted to be statistically important for cluster membership are included in the table, supplemented by the *hg* index, which, in previous studies, is suggested as a robust indicator for the rank comparison and division of scholars (Wildgaard, 2015b); (Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2009). The negative sign reveals the direction of the effect, indicating that the effect is bigger for the second cluster.

The biggest effect sizes were observed when Clusters 1 and 2 were compared to Cluster 4. Across all disciplines, Cluster 1, presented on hybrid indicators, differed greatly in regards to the other three clusters since the effect sizes are very large. Noticeably the effect sizes between Clusters 2 and 3 were the smallest, but still generally benchmarked as having a medium to strong effect. Publications and citations displayed very large discriminatory power between clusters. Consequently, *CPP* had a large to very large practical effect for cluster membership, which increased when the clusters of lower performing researchers (Cluster 1 and 2) were compared to the high performers (Clusters 3 and 4). The effect of academic age was

Table 8
Percent within seniority within the 4 clusters and odds ratio associated with each seniority.

Astronomy	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	% within seniority	Odds	% within seniority	Odds	% within seniority	Odds	% within seniority	Odds
PhD	80	13	–	–	20	0.5	–	–
Post Doc	60	7.6	12	0.2	27	0.7	–	–
Assis Prof	30	1.1	27	0.8	42	1.6	–	–
Assoc Prof	5.3	0.07	46	3.4	33	1.0	15.8	2.9
Prof	–	–	37	1.4	37	1.2	25	4
Environmental Science								
PhD	100	–	–	–	–	–	–	–
Post Doc	82	12	17	0.2	–	–	–	–
Assis Prof	46	2	48	1.2	5	0.1	–	–
Assoc Prof	24	0.5	51	1.7	23	1	–	–
Prof	13	0.2	39	0.7	45	4.5	2	–
Philosophy								
PhD	87.5	4.9	12.5	0.2	–	–	–	–
Post Doc	72.7	1.9	22.7	0.5	4.5	0.6	–	–
Assis Prof	65.9	1.4	27.2	0.7	4.5	0.6	2.3	1.3
Assoc Prof	65.8	1.4	26	0.6	8.2	1.5	–	–
Prof	42.7	0.3	46.7	2.6	6.7	1.0	4.0	6.0
Public Health								
PhD	88.9	11.2	11.1	0.2	–	–	–	–
Post Doc	85.7	9.0	14.3	0.2	–	–	–	–
Assis Prof	60	2.2	30	0.7	10	0.5	–	–
Assoc Prof	30	0.3	52	2.9	16	1.0	2.0	0.5
Prof	20	0.2	34.5	0.9	34.5	4.4	10.3	11.7

large, especially between Cluster 1 and the other three clusters ($g > 1.1$). Average seniority scores in the clusters differed by $g \leq 2.3$ standard deviations, and displayed that seniority had a trivial to very large practical effect on cluster membership dependent on discipline and cluster pair, the smallest effect in Philosophy and the largest in Astronomy and Public Health. Gender scores were consistent, displaying a marginal difference between *g*0 and *g*0.4 standard deviations. Collaboration displayed overall a large to very large effect in Astronomy and Philosophy, and a medium effect in Environmental Science and Public Health. The hybrid indicators had the strongest discriminatory effect within disciplines and between pairs of clusters. Specifically, the *hg* indicator had a very large effect across all four disciplines and all cluster combinations ($g > 1.1$). In Astronomy, Philosophy and Public Health the indicators predicted to be of statistical importance for cluster membership, respectively the *h2, Q2* and *e* indicators, did indeed have the largest effect within and between clusters, but the difference to *hg* or the other hybrid indicators was marginal. In Environmental Science the *hg* indicator and other hybrid indicators provided greater discriminatory power than the statistically important indicator *sum pp top prop.* Further tests are needed to investigate if using the *hg* indicator across all disciplines would result in satisfactory clusters and eliminate the challenge for defining disciplinary specific indicators.

4.2.3. Likelihood of researcher placement in cluster

In researcher assessment, seniority of the researcher is a typical consideration for contextualizing bibliometric statistics. To explore if Cluster 1, the cluster with the lowest publication count, was dominated by junior researchers, and consequently if Clusters 2, 3, and 4, the clusters were publication count increased, were progressively dominated by more senior researchers, we computed the odds of a researcher who belongs to a particular seniority belonging to Cluster 1, 2, 3 or 4. SPSS crosstabs and risk statistics were used for this analysis. Table 8 presents the likelihood for seniority classes belonging to a cluster and is interpreted accordingly: An odds ratio of 1 means that the seniority has a similar likelihood to other seniorities to belonging to a cluster. The larger the odds ratio the more likely this event is expected to occur, odds ratio less than 1 are interpreted to indicate a protective effect, meaning that the seniority is less likely to be in this cluster. An odds ratio cannot be used to assess the causation of seniority and the resulting clustering, but it can help us determine if seniority influences the composition of clusters.

The results show that junior researchers (PhD students and Post Docs) were more likely to be in Cluster 1, the low performers, the likelihood of Assistant and Associate Professors being placed in specific clusters was unclear, and Professors were most likely to be in Cluster 3 or 4, the high to extremely high performers. In Astronomy, PhD students were 13 times more likely than other seniorities to be grouped in Cluster 1 and the odds of a Professor being grouped in Cluster 4, were 4 times as likely than compared to the other seniorities. In Environmental Science, post-doctoral students were 12 times as likely to be in Cluster 1 and Professors 4 times as likely to be in the high scoring Cluster 3. Similarly in Philosophy and Public Health, PhD students were respectively 5 and 11 times as likely to be in Cluster 1 and Professors 6 and 11.7 times as likely to be in Cluster 4. The observed odds ratios could be inaccurate because of selection bias, as our sample of researchers is by no means random. Further, seniority can be dependent on other factors, such as the promotion policy of the institution the

Table 9
Cluster composition. Discipline, academic age in years, publications, and share of researchers by academic seniority.

Discipline	Age (min;max)	Publications (min;max)	Seniority
Astronomy			
Cluster 1	7 (2;18)	15.9 (2;53)	12/15 PhD, 29/48 Post Doc, 8/26 Assis. Prof, 4/66 Assoc. Prof
Cluster 2	21 (10;33)	53.4 (15;139)	6/48 Post Doc, 7/26 Assis. Prof, 31/66 Assoc. Prof, 14/37 Professors
Cluster 3	15 (3;34)	70.4(5;171)	3/15 PhD, 13/48 Post Doc, 11/26 Assis. Prof, 20/66 Assoc. Prof, 15/37 Professors
Cluster 4	23 (11;33)	213.2(93;327)	11/66 Assoc. Prof, 8/37 Professors
Environmental Science			
Cluster 1	9 (2;31)	10.6 (21;180)	3/3 PhD Students, 14/17 Post Doc, 18/39 Assis. Prof, 21/85 Assoc. Prof, 7/51 Professors
Cluster 2	17 (7;36)	37.6(425;425)	3/17 Post Doc, 19/39 Assis. Prof, 44/85 Assoc. Prof, 20/51 Professors
Cluster 3	22 (9;34)	77.3(6;102)	2/39 Assis. Prof, 20/85 Assoc. Prof, 23/51 Professors
Cluster 4	32	425(1;26)	1 Professor
Philosophy			
Cluster 1	9 (0;33)	6.4(1;28)	7/8 PhD, 16/22 Post Docs, 29/43 Assis. Prof, 48/74 Assoc. Prof, 32/75 Professors
Cluster 2	15 (3;33)	24.1(3;140)	1/8 PhD, 5/22 Post Doc, 12/43 Assis. Prof, 19/74 Assoc. Prof, 35/75 Professors
Cluster 3	22 (15;30)	35.9(61;119)	1/22 Post Doc, 2/43 Assis. Prof, 6/74 Assoc. Prof, 5/75 Professors
Cluster 4	18 (7;33)	99.7(10;64)	1/43 Assis. Prof, 3/75 Professors
Public Health			
Cluster 1	9 (0;25)	17.3(1;56)	8/9 PhD, 12/14 Post Doc, 18/30 Assis. Prof, 15/50 Assoc. Prof, 6/29 Professor
Cluster 2	15 (4;34)	48.2(7;146)	1/9 PhD, 2/14 Post Doc, 9/30 Assis. Prof, 26/50 Assoc. Prof, 10/29 Professor
Cluster 3	21 (10;33)	140.7(62;288)	3/30 Assis. Prof, 8/50 Assoc. Prof, 10/29 Professor
Cluster 4	22 (19;28)	249.7(18;661)	1/50 Assoc. Prof, 3/29 Professors

researcher is affiliated to, information that we do not have. The number of publications could be a positive confounder in this analysis and the effect of academic age could positively or negatively modify cluster placement. Subsequently, Table 9 presents descriptive statistics as a supplement to the odds analysis.

4.2.4. .Changes in within cluster rankings

Even if researchers belonged to the same cluster, they were ranked within the cluster in a different order, depending on the indicator. The following projections are illustrated using Astronomy researchers, because this disciplinary group provides a comparable amount of researchers in each cluster.

Fig. 2, top left, shows cluster analysis projection on the Fractionalized citations, *Fc*, and Fractionalized publications, *Fp*, axes. Citation and publication count generally increased from Cluster 1 through 4. However, large differences in the ranked performance of the researchers were observed. For example, researcher ID164, Cluster 4, ranked in first place using *Fp* (fractional publication count) whereas ID162, Cluster 4, ranked 82nd out of all of the 192 Astronomy researchers. ID183, Cluster 4, ranked first place using *Fc* (fractional citation count) while the other 18 researchers in Cluster 4 were spread between 2nd to 65th ranked position. However, when all the Astronomy researchers were ranked after *C*, the raw citation count, the 19 researchers in Cluster 4 were ranked in the top 19 positions.

Fig. 2, top right, shows the projection on the *C* and *Fc* axes. Whereas *C* favours researchers with a high total citation count, normalizing citations to the number of contributing authors, *Fc,* reduced the dominance of Cluster 4 researchers in the top rank positions, lifting researchers from the other clusters up the rankings. Generally, researchers who collaborated with $\leq$3 partners moved up the rank, while those who collaborated with $\geq$4 authors fell in rank position. Of course, the amount of papers and citations the collaboration is distributed across further influenced this trend.

Fig. 2, middle left, shows the projection on the citation count, *C*, and h-index, *h*, axes. The distribution is linear and there are clear thresholds between the clusters. Within the clusters, the top researchers in Cluster 4 and bottom researchers in Cluster 1 were ranked in the same positions on both indicators, the rank changing by roughly $\pm$1 rank position. The ratio *P* to *h* (total publications divided by *h* value) determined rank position. If a researcher had a ratio *P:h* of $\geq$3 they fell in rank position, and if the ratio was <3 they gained in rank position.

Fig. 2, middle right, shows the projection of the *m* and *A*-index. These two indicators are interesting to compare as *m* takes the median of citations to articles in the *h* core and *A* takes the arithmetic mean. The results show *m* rewarded researchers with higher rank placements than *A*. The researchers in Cluster 3 and 4 were placed about 6 positions higher on *m* compared to *A*, however Clusters 1 and 2 were raised up to 20 positions higher on *m.* A dependence ratio similar to *P:h* was found on the projection of the *m* and *A* axes.

Fig. 2, bottom left and bottom right, illustrates indicators that promoted researchers in Cluster 1, 2 and 3 up the rankings. The projection, bottom left, shows the *CPP* and *hnorm* axes. Researchers in Cluster 1 scored the lowest citations per paper, *CPP*, see Table 3, however it was these researchers, that scored some of the highest values on *hnorm*, which like *CPP* is a method of computing citations per paper. Fig. 2, bottom right, shows projections of the *h* and *m-quotient* axes. The *m-quotient* normalizes *h* for the academic age of the researcher. Ranking using the *m-quotient* promoted researcher ID4 (Cluster 3) from *h* rank position 104 to *m-quotient* rank position 1. This researcher had 25 publications, 529 citations, *h* 15 and an academic age of 3 years. Similarly, ID 23 (Cluster 1) also had *h*15 and rose from *h* rank position 100 to *m-quotient* rank position 35. This researcher had a career of 8 years, 53 papers and 634 citations. Conversely, researcher ID114, *h*15 Cluster 2, fell in rank

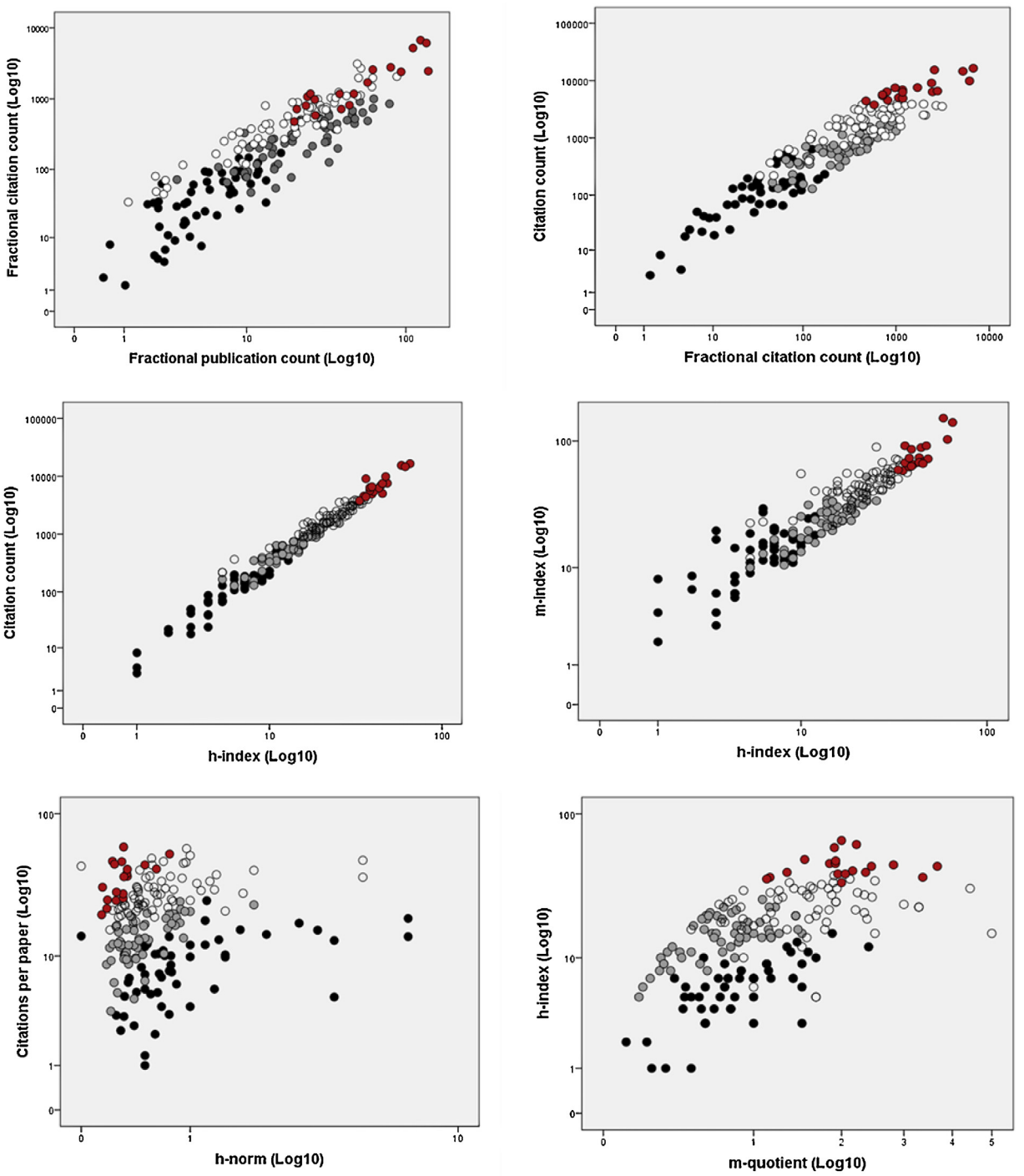

Fig. 2. Clustering in Astronomy. Projection on bibliometric indices axes.Legend: Cluster 1 = black, Cluster 2 = grey, Cluster 3 = white, Cluster 4 = red.

position from *h* rank 103 to rank position 153 using the *m-quotient.* ID114 had an academic age of 23 years, 50 papers and 575 citations. In this indicator, a large academic age value appeared to cause a fall in rank position.

The above analyses were repeated for Environmental Science, Philosophy and Public Health, apart from Cluster 4, which in all three disciplines was very small, ≤4 members. The distribution of researchers within Clusters 1, 2, and 3 and patterns of change in the rank position were the same as observed in Astronomy. The observed ratio *P:h,* which determined a rise or fall in rank position, was also present in all disciplines.

5. Discussion and conclusion

Author-level indicators are becoming institutionalized in national and university assessments. It is important the bibliometric community explore the disciplinary and seniority appropriateness of such indicators, and come with informed recommendations for their application. In this paper, the Two-Step cluster analysis was explored through a 7-stage process, as a method to identify such disciplinary and seniority appropriate indicators. Like other clustering methods, our approach can be criticized for producing arbitrary cluster solutions, as it is difficult to use bibliometric indicator values as a base on which to form distinct, unambiguous clusters. However, as argued in the Choice of Clustering

Method: section (Stage 4), the Two-Step cluster approach was the logical choice of technique. It made sense for our dataset, the objectives of the analysis and was rationally useful for the task (Schneider & Borlund, 2007a).

5.1. Stage 7: informed interpretation of clusters

5.1.1. The two-Step cluster analysis

The accuracy of the clusters is dependent on the thoroughness of the preliminary data preparation processes, also discussed in (Chawla, 2006); (Su et al., 2013) hence in this paper preparation was extremely thorough (Stages 1–3). Yet our data collection, even though it was thorough, was limited to WoS indexed papers. It is important to make clear here at the start of the discussion, that using Scopus or Google Scholar, as Supplementary data-sources, could have identified important output on the researcher's CV not present in WoS. This would have resulted in entirely different bibliometric profiles of the researchers included in this study and subsequently an entirely different cluster solution. The potential limitations of WoS in evaluation is addressed further in the Limitations section of this paper. Yet bear in mind, it has never been the intention of our study to definitively group researchers and evaluate their prestige, but to investigate the appropriateness of clustering as a methodology to identify indicators for application in research evaluation exercises.

5.1.2. Cluster definition

Based on the WoS data, the clustering algorithm identified 4 groups of researchers that were substantively interpreted as low, middle, high and extremely high performers. As performance indices behave differently on different data collected from different sources, the interpretation of an optimal number of clusters is arbitrary without quality control of the data and strong methodological arguments (Liu et al., 2009). Our confidence in the method must not override validating the proposed cluster solution with common sense observations of 1) the completeness of the data, 2) what the data represents, and 3) the variables that make up the dataset. We chose not to transform the data, well aware that normalizing the data in this way ensures that the chosen distance measure gives equal weight to each variable. Without normalization, the variable with the largest scale will dominate the distance measure and cluster composistion (Gower, 1967). Whether it is desirable to normalize is domain specific and dataset specific. The question regarding our dataset is, if it is defensible to normalize carte-blanche across four very different disciplines and if the dominance of the big values dominating the clustering solution is problematic? Statistics alone do not lead to a definitive decision on the number or composition of clusters and more than one solution can be appropriate. Similarly, Goodness of Fit and likelihood statistics are a *guide* to selecting the number of clusters and should not be adhered to automatically. We found that interpretation of within and between cluster similarities and dissimilarities, based on statistics alone, was uninformative and had to be supported by other

methods: Therefore, the approach recommended in this paper is not to rely entirely on statistical clustering parameters but to verify cluster composition and interpret indicator scores against disciplinary characteristics and the researcher's CV. Only in combination with this knowledge, can statistics be used to inform clusters and select the most appropriate indicators.

Importantly, it is uncertain if the cluster analysis resulted in the *correct* division of researchers. We do not know the ground truth, though the applied clustering does appear to provide a defensible solution. The cluster analysis identified indicators that produced clusters that we can argue appeal to certain disciplinary characteristics (the *h2, sum pp top prop, Q2* and *e* indicators). Clusters 1 and 4 worked well for identifying the extreme cases in the dataset, separating the junior from the senior researchers. However, Clusters 2 and 3 were very muddled. In Astronomy in particular, the number of publications and citations had a weaker effect on cluster membership than collaboration, which had a very large effect on cluster membership in these clusters, Table 7. In Environmental Science, Philosophy and Public Health the reception of the researcher's work measured in citations was a characterizing factor for cluster membership – having a stronger effect than the number of publications. Looking at the demographic variables, neither gender nor seniority defined the clusters. Other demographic factors that could distinguish clusters satisfactorily, but not investigated in this paper, are the nationality of the researcher, the prestige of the university and the academic culture these variables brings with them.

5.1.3. Likelihood of researcher placement in cluster

PhD students and Post Docs had the greatest odds of being placed in Cluster 1 and Professors in Cluster 4. The odds placement of Assistant Professors and Associate Professors was uninformative. This is not a failure of our study. It shows that bibliometric values transverse seniority classifications, suggesting that seniority titles are not a key influence in indicator design or stable parameter for cluster division. Rather seniority is just one principle for interpreting the indicator values, within a conceptual framework that provides other principles, assumptions and rules that holds together ideas comprising broader concepts of how bibliometrics can inform us about disciplinary research activities and academic structures. Within

this framework, productivity and citations are traditionally assumed to be strongly correlated to seniority (Costas et al., 2010), but they are just as likely to be linked to other factors such as gender, nationality, specialty, and timeliness, etc. Academic age for example was statistically significant for cluster placement with a substantial increase in the odds of researchers with a higher academic age being placed in Cluster 3 and 4 (15% increase with each unit increase in age). However, this did not explain all of the variance in cluster placement, and the effect analysis showed the strong influence of the hybrid indicators that predominately distinguished the clusters and at closer inspection, steered the rank placement of researcher's within the clusters, section 4.2.4.

5.1.4. Stability, validity and heterogeneity of the clusters

Researchers belonging to Clusters 1 and 4 were bibliometrically very different from each other, while researchers in Clusters 2 and 3 were bibliometrically similar. Studies before this one have commented that bibliometric indicators are most useful in identifying differences between top and bottom researchers rather than the middle set where the researchers are bunched tightly together, i.a. (Frey & Rost, 2010;; Meho & Yang, 2007). The differences have been discussed specifically in the light of the researcher's academic seniority (Costas et al., 2010) and between and/or within academic domains (Claro & Costa, 2011); (Archambault & Gagné, 2004). Therefore, the observed division between the clusters in our study was expected. Our substantive interpretation of the clusters as low, middle, high or extremely high performers was deceivingly simple, and we found that the interoperability criteria for a good classification were not fulfilled because of the similarities between Clusters 2 & 3, Bacher et al., 2010. However, the clusters displayed stability, removing a small number of cases (though not the outliers) did not change the resulting clusters, and statistically the clusters displayed relative validity, as the classification was better than the null model which assumed no clusters were present in the data – hence the usefulness of this relative validity statistic is highly questionable. Likewise, further validation tests in respect to the similarity of the clusters are also questionable, because the first step of the Two-Step clustering algorithm identifies pre-cluster groupings based on a hierarchical method of partitioning. Common measures such as the Rand Index are only valid for strict partitionings. The cut-off point between clusters of researchers based on the ground truth would first have to be determined, which we do not have. The Two-Step cluster belongs to a family of exploratory data analysis techniques called unsupervised learning, where there is no error or reward signal to evaluate potential solutions as there is no way to evaluate the stability of the clustering algorithm. There could be a bias in the SPSS clustering algorithm towards partitions that are in accordance with a certain clustering criterion, and a different algorithm in another statistical program, like SAS or R, may have provided different solutions. The heterogeneity between clusters was tested by comparing cluster averages (mean) per indicator, and in the majority of cases, there was a statistical difference between clusters, but this is not necessarily interpreted as an *important* difference between the four groups of researchers. The advantage of using mean cluster scores is that they are less susceptible to noise and outliers. A median score, if available, might have provided an even more superior solution.

5.1.5. The statistical importance of each indicator as a predictor of cluster membership

Different indicators were statistically identified as stronger predictors of cluster membership and within cluster rank in the different disciplines: in Astronomy the *h2* indicator, Environmental Science *sum pp top prop,* Philosophy *Q2* and Public Health *e*. These indicators are designed to capture different aspects of researcher performance: cumulative achievement (*h2*), papers at the top of the field (*sum pp top prop)*, effect of all productive papers (*Q2*), and, production and effect of highly cited papers (*e*). Each indicator is calculated using different mathematical formulae that, it can be argued, better fit the publication and citation characteristics of each discipline, e.g. *h2* corrects for the ratio many papers to few extremely highly cited papers that is a common characteristic of output in Astronomy and thus produces more granular and comparable rankings of researcher productivity and impact. Previous studies have shown that it does not make sense to compare researcher rankings using the same indicators across different disciplines, and our observations support the discussion that some indicators are more appropriate in some disciplines and for some seniorities than others (Costas et al., 2010); (Díaz-Faes et al., 2015). Nevertheless, identifying substantial disciplinary and seniority indicators is difficult and fraught with caveats. Hence, there is a great interest in producing universal indicators that do not favour one discipline over another and are specifically designed for valid cross-disciplinary comparisons (Claro & Costa, 2011; Sidiropoulos, Katsaros, & Manopoulos, 2007; Vinkler, 1996). The effect analysis, discussed further in the sub-section 5.1.7, revealed the ranking indicator *hg* as an important universal approach for demarcating researchers, outperforming in fact the statistically significant indicators in defining clusters within disciplines. The *hg* indicator is designed to produce granular ranks of researchers, which it does very well in our clustering model. Yet multiplying the *h* and *g* indicators together and taking the square root of the sum, as is the construction of *hg*, has no direct meaning in terms of papers and citations of a researcher and can lead to hasty judgements about performance if this indicator is used alone (Franceschini & Maisano, 2011). The effectiveness of a particular indicator for cluster membership can only truly be interpreted in relation to other, similar studies (Glass, McGaw, & Smith, 1981). Glass et al. suggest that combining effect sizes from multiple studies into a single effect size in meta-analyses would indeed help us to understand which factors or characteristics of the sample group, and which indicators, account for differences in researcher performance. The practical importance of the effects, just like the statistical importance, depends entirely on its relative "costs and benefits". Therefore, we maintain that the indicators, deemed *statistically* important in defining the clusters, are not necessarily the same as the indicators deemed *practically* important in defining clusters, and we can only identify the *scientifically* important indicators with knowledge of the data and knowledge of whom we are evaluating bibliometrically.

Continuing our case in point, seduced by the success of our statistical model, we easily found patterns in disciplinary characteristics that defended the appropriateness of using the statistically important indicators, *h2, sum pp top prop, Q2* and *e,* to define clusters. Yet, what if these indicators mathematically produced a better fit to the *F* statistic (which was the test used to predict the statistical importance of the indicator for cluster membership) and not a better fit to the performance of researchers? Ranking researchers by the predicted indicators of importance, Table 6, explored this consideration. It appeared that the sole importance of these indicators was to produce clear thresholds between groups of researchers in the cluster algorithm and thus demarcate the four clusters. That is, ranking the researchers using the predicted indicators of importance grouped the researchers into four definite, separate clusters, $F = 1$, whereas ranking the researchers with indicators with weak prediction strength, $F < 1$, produced muddled, overlapping groupings. Hence, based on the division of researcher rankings using the *F* statistic, as low, middle, high or extremely high performers, *statistically* significant different clusters were produced. But then this not the same as the *F* statistic produced *scientifically* important clusters. In the cluster solution we interpreted the success of the computer model and interjected our own biases and expectations about research performance in interpreting the worth of the clusters. The clustering algorithm fueled these assumptions, as it used different indicators in different disciplines to group the researchers. Why was this? Here we return to the dominace of the variable with the largest scale, Section 5.1.2, which has indeed proved problematic for the definition of scientifically relevant clusters and contributed to the false assumptions we drew in interpretting the clustering model. The distances between variables changed dramatically in our data, we changed for example from the *m*-quotient with values ranging between 0.2 to 3.6, and *C* with values ranging between 1 and 16,481. Normalizing the variables would have put all the variables into the same range, the variables would have been weighted equally. Appropriate nomalization for each discipline, perhaps for each indicator, must be incorporated in our methodological process to help us identify clusters not affected by the scale between variables, but as yet we have not determined the appropriate method of normalization.

Understanding now the limitations of our statistical model, can we conclude *anything* about how to identify appropriate indicators of academic performance or are we limited to interpreting the mathematical performance of the clustering algorithm?

5.1.6. Composition of the clusters

Although the clustering algorithm differentiates between researchers based mainly on different statistical properties, interesting scientific observations can still be derived when considering cluster composition based on the dominate indicators. Table 6, illustrates heterogeneity of the clusters for a small proportion of the 44 indicators, between 16 and 31 dependent on the discipline.

Researchers placed in Cluster 1 were characterized by; citing themselves the most, having the highest proportion of un-cited publications, gaining benefit for increased rank placements (currency of their work: productive papers they have in their portfolio). The results in Cluster 1 is likely attributed a dominance of junior researchers with a small amount of publications that are 'young', rather than specific publication and citation patterns of the research discipline. Cluster 1 researchers, dependent on discipline, had an average academic age between 7 and 9 years, and a title of PhD student or Post Doc researcher. Whereas, Cluster 2 researchers had an academic age between 15 and 21 years and were Assistant, Associate or full Professors. Cluster 3 researchers also had an academic age of 15–21 years but were predominately Associate Professors or Professors. Researchers in Cluster 4 were had an academic age of 18–23 years and were Associate Professors or Professors. In spite of similarities in academic age and academic titles, production and citation profiles between clusters was vastly different. Researchers (Cluster 1–4, publication range with citation range in parenthesis) had, 6–17 (5.6-150), 24–53 (65.5-741), 35–140 (432.7-2602.9) and 99–425 (1971.2-14,141) publications and (citations). No statistical difference between clusters were found including indicators of production *(P, Fp),* indicators that count collaboration and cognitivity (no adjusting for age), these being the indicators *APP, mean pp collab, mean pp int collab, POPh*, and indicators that normalize the number of citations across all publications, *CPP, Fc, FracCPP.* Interestingly, the researchers in Clusters 3 and 4 did not collaborate any more on average than researchers in the other two clusters (*APP*), but they did display higher scores on the indicators of inter-institutional collaboration (*mean pp collab*) suggesting a more diverse co-authorship network than Clusters 1 and 2 researchers. Paradoxically, Cluster 3 and 4 researchers also scored highest on the indicator that rewarded independence, *POP h*.

In regards to the "prestige" indicators, we observed that Cluster 1 and 2 researchers were cited 10%–40% less than the expected field average according to WoS criteria; *mncs* (adjusts for field, publication type and publication year). In contrast, Cluster 3 and 4 researchers were cited between 40%–90% more than expected. For journals in which Cluster 3 academics published a higher normalized average citation score of the journals was found compared to Cluster 2 researchers (assessed by indicators of prestige, *mcs, mncs, and max mjs mcs*). Yet on a lower threshold for defining impact within specialty rather than field, $T > ca$, Cluster 3 and 4 researchers were not necessarily cited more than expected. Judging by their low scores on *mncs* and $T > ca$, and high scores on *NprodP* researchers in Cluster 2 produce papers that performed better than average for articles in sources important for their specialty rather than producing the broader field impact of Cluster 3 and 4 researchers.

According to the *%sc* and *%nnc* indicators researchers in Clusters 3 and 4 had the fewest non-cited documents. This makes sense as Clusters 1 and Cluster 2 researchers had the fewest and newest papers, and highest proportion of most recent citations for all publications (*PI and Cless5*). The exception was Philosophy, where researchers scored the lowest *%sc* compared to the other disciplines, although with the *%sc* still increasing from Cluster 1 through Cluster 4 researchers. This was likely a methodological abnormality due to restricted data collection from WoS, as WoS, for Philosophy, appears to have

insufficient coverage. Hence, philosophy researchers scored the lowest on the effect and prestige indicators, Appendix A. We suspect these scores give a distorted view of researcher performance due to the lack of coverage in WoS but they could further be affected by disciplinary citation practice, in that citations in Philosophy do not appear to be given as readily in comparison to the other disciplines. The effect analysis showed citations as an important variable in cluster composition, therefore controlling or normalizing for the value of citations within a disciplinary discourse is vital for the strength of the clusters and for wise cross-disciplinary comparisons based on bibliometric indicators in the context of science policy (Podlubny, 2004; Crespo, Li, & Ruiz-Castillo, 2012).

In general Cluster 3 and Cluster 4 researchers scored highly on the *h*-dependent indicators of impact, but also on some *h*-independent indicators of impact (*g* and *AW* indices), that normalize for high and/or low cited publications. Seemingly, the majority of indicators favour senior researchers with a considerable number of publications and hence citations. Accordingly, the same indicators are not as informative for researchers with a smaller catalogue of publications, typically junior researchers. On average, the hybrid indicator scores doubled between clusters, distinguishing cluster boundaries by ranking researchers more granularly, but because of their sensitivity to changes in the number of publications, citations and time they are quite complicated and might not reflect real person "value".

5.1.7. *Effect size or why the F statistic was not enough*

The statistical significance of an indicator in cluster prediction depends on two things; the size of the effect and the size of the sample. One would get significant results either if the effect were very big (despite having a small sample) or if the sample were very big and the effect size was tiny. Statistical significance does not tell us the size of the effect, i.e. the practical importance of the indicator. Therefore, we chose to report the effect size. Effect size quantifies the size of the difference between two groups, and is indicative of a true measure of the significance of the difference (Schneider, 2013; Coe, 2002). Cohens classification of effect sizes (small, medium, large and very large) provide a general guide to interpreting the effect of the indicator in clustering solutions, but these need to be informed by context and they can only be argued as informative if the objects we are comparing are similar (apples and apples, not apples and oranges). One feature of effect sizes is that they can be converted into statements about the overlap between clusters in terms of a comparison of percentiles and rank order (Coe, 2002). For example, following Coe's argument, an effect size of 0.7 means the average academic age of a researcher in Astronomy, Cluster 3, is 0.7 standard deviations above the average researcher in Cluster 2, and exceeds therefore the academic age of 76% of the researchers in Cluster 2. With these two similar sized clusters, we can further estimate that the average researcher in Cluster 3 (i.e. ranked 31st within the cluster on Academic Age) would have been placed the 6th highest in Cluster 2. Another example is the average effect size of −0.5 between Cluster 2 and 3 publications, again for researchers in Astronomy. This tells us that the average score of the researcher in Cluster 3 is 0.5 standard deviations above that of the average researcher in Cluster 2, and hence exceeds the scores of 69% of the Cluster 2 researchers. Whereas an effect size of 1.8 (comparing publications between Cluster 1 and Cluster 3 in Astronomy) would raise the average researcher in Cluster 3 to be, level with or exceed the topped ranked individual in Cluster 1. Conceptualizing the effect sizes between clusters in this way gives us a palpable understanding of what the standard deviations actually tell us about cluster membership and the real difference between Clusters based on performance indicators. Yet we have to be cautious, in that interpreting effect sizes depends on the assumption of a normal distribution, and interpreting effect sizes in terms of percentiles is very sensitive to violations of normality. It may be difficult to make a fair comparison between clusters in bibliometrics without transforming the data, as the data we are working with is non-normal distributed and the clusters and researchers within the clusters have different average values and standard deviations, and overlapping confidence intervals. We need to consider the effect of the underlying distribution in interpreting the results, they may be trivial or they may be important for the stability of our clusters (Schneider & van Leeuwen, 2014). We need to do more specialized analyses that investigate differences in the distributions of author profiles between the four clusters and if normalizing for these differences harms our perception of a researcher's performance. This is not within the scope of this article.

Our study has revealed several critical concerns that should now be further investigated relative to the application of statistical procedures in research evaluation. Firstly, the inadequacy of limiting the bibliometric data to journal articles indexed in WoS as a proxy for the academic output and impact of researchers. This directly influenced the bibliometric profile of our sample of researchers, the indicator values, the cluster solution, the composition of the clusters and the conclusions we drew on the performance of researchers in these clusters. Yet the production of journal articles and citation counts registered in the WoS index remains the staple for bibliometric analysis because it is here we readily find documentation of the presentation, discussion, criticism and validation of research. Importantly the WoS database structure supports reproducibility of methods used to collect bibliometric data, unlike Google Scholar, which although presents a broader selection of academic output beneficial for national and local production, does not support repeatable identification and extraction of bibliometric data. Therefore, grouping researchers based on articles and citations identified in WoS, although limited, is interesting as it robustly indicates the participation of the individual researcher in a formally documented and verifiable communication process. Researchers whose dissemination was not covered in WoS were at a bibliometric disadvantage in our study – simply because output that was not counted became invisible. Nevertheless, it is not only the sources in which we gather our bibliometric data that need to be assessed as part of statistical methodologies. The quantitative procedures for research evaluation also need to be subject to ongoing assessment. They should be transparent at all stages of the procedure and hence open to discussion and modification to reflect the changing objectives of the mission of the evaluation and importantly to improve measurement techniques. It is useful to suggest techniques, like the 7-stage method in this paper, to highlight the technical

issues of producing reliable statistical analyses and recommending reliable bibliometric indicators. There are a variety of ways that this can be achieved, such as improving the quality of the data used to calculate the indicators which entails investing in the time and cost it takes to produce quality data: collecting all relevant publications, sourced from multiple databases, with duplicate citations removed. We could argue for alternative clustering methods based on the character of our data, specifically our next step is to investigate how different normalization approaches affect our cluster solution and perception of researcher performance. We could also stratify the researchers within the clusters, to see if there are citation advantages given to researchers who are socially closer prestigious institutions and likewise stratify for researchers from research heavy and teaching heavy institutions to understand more about the depth and breadth of institutional policies on the profiles of researchers in the different clusters. If we can in bibliometric evaluation show through robust cluster analyses, that making a small and inexpensive change in for example institutional policy, would raise academic production or impact by an effect size of even as little as 0.1, then this could be a very significant improvement. Particularly if the improvement applied uniformly to all researchers and even more so if the effect were cumulative over time.

The 7-stage cluster methodology proposed in this paper is an effective first draft in putting into context the issues that surround statistical analyses of researcher performance and the potential methods have for supporting the researcher in evaluation. It offers a practical context for developing robust, transparent methodologies. The statistical methodology must however be supported by qualitative data, e.g. researcher's CV, as well as knowledge of disciplinary publication and citation practices to ensure we are evaluating the "value" of the person and not the value of the statistical model.

5.2. Limitations

Only papers registered in WoS were used to assess researcher performance. The amount of publications covered in WoS was fractional compared to the amount of academic publications listed on the researchers' CVs, thus the indicators and resulting clusters are a fractional representation of the production and effect of the researchers' work. The Philosophers listed in total 14,762 publications (articles, reviews, conference papers) on their CVs and we were able to identify 3753 of these publications in WoS, approximately 24% of their papers. Likewise, we identified 80% of the papers listed by the researchers on their CVs in Public Health, 58% in Astronomy and 47% in Environmental Science.

The cited references function in WoS was not used, meaning that citations in books, proceedings and other sources not indexed in WoS were missing from the analysis.

The effect of coverage of researchers and disciplines in citation databases must not be ignored in the interpretation of bibliometric studies of researcher performance. Lack of coverage in WoS can result in distorted indicator values that do not fully represent the researcher's publication activity and impact.

Striving to identify all a researcher's publications by searching multiple databases is simple enough, as is setting up criteria defending the inclusion or exclusion of particular types of publications depending on the mission of the evaluation. However, removing duplicate citations to publications indexed in multiple sources is labour intensive. Yet this may be a necessary to improve the accuracy of performance indicators.

The obvious confounder associated with cluster membership, is the academic age of the researcher. In this paper academic age is calculated as the number of years from 2013 since the researcher's first publication recorded in WoS, not the actual number of years they have been active as a researcher. Academic age is thus highly dependent on coverage of the researcher in WoS, and coverage can distort the magnitude of bibliometric indicators that adjust for the career length of the researcher.

6. Conclusion

The 7-stage methodology proposed in this paper is used to explore performance-based clustering of researchers, with the aim to identify disciplinary and seniority appropriate author-level indicators. Stage 1) data-collection, 2) description of data, 3) presentation, calculation and statistical description of bibliometric indicators, 4) a rationalized choice and application of the cluster algorithm and clustering statistics, 5) presentation and statistical description of clusters, 6) tests of the stability and strength of the clusters, and finally, 7) informed interpretation of the clusters. Using these stages added clarity, and thus we were able to confidently challenge the robustness of the methodology and question if the resulting clusters and indicator values amount in convincing evidence of researcher performance. No seniority appropriate indicators were identified. The clustering method did however identify different indicators in the different disciplines as more statistically important for clustering similar researchers and demarcating dissimilar researchers, grouping and ranking them as low, middle, high and extremely high performers. The practical importance of these indicators was explored in an analysis of effect size, where alternative indicators were found more than or just as influential in cluster membership as the statistically important ones. Gender did not have a strong effect on cluster membership, and while the effect of academic seniority on cluster membership was present, stronger influences such as the academic age of the researcher, were identified. Unsurprisingly, the main conclusion of this study is that applying statistical methods to evaluate researcher performance is complicated. The strength of our methodology is that it makes clear that in evaluation of individual researchers, statistics cannot stand alone. It is vital in the application and interpretation of cluster analysis we do not get caught up in the statistical significance of our results and forget to judge if the results are at all important (Bach, 2011; IEEE, 2013; Aksnes, 2009). We are aware that all ordinations are wrong to some extent, and introduce bias into the cluster solution. Another cluster methodology or statistical software could have produced another cluster solution and cluster composition. The wise and unwise use of statistics in

bibliometric evaluative studies has been addressed predominately by (Schneider, 2013), who recommends an overview of the challenges in (Kline, 2013). Even though it is unwise to generalize the results of this study outside of our dataset, it is important to do studies like this one, which critically investigate the usefulness of statistical models and the interpretation of bibliometric indicators. This will help us as a community learn more about the advantages and disadvantages of quantitative analyses of researcher performance, and help us illuminate the appropriate and inappropriate application of statistical.

Acknowledgements

This work was partially funded by ACUMEN (Academic Careers Understood through Measurement and Norms), FP7 European Commission 7th Framework "Capacities, Science in Society", grant Agreement: 266632. Opinions and suggestions contained in this article are solely the authors and do not necessarily reflect those of the ACUMEN collaboration. I would like to thank the reviewers for their positive reception towards publishing the negative findings presented in this paper and their very useful suggestions for improvement.

Appendix A. 44 bibliometric indicators of individual performance. The columns, from left to right, present the full name of the indicator, the abbreviation, the definition and the aim of the indicator, as proposed by the creator of the indicator. Categories correspond to Stage 3a (described in text).

Dimension		Indicator	Abbr.	Definition	Aim to Assess
Publication- Based		Number of publications	P	Total number of publications by the researcher	Production
		Fractional publications	Fp	Each publication divided by number of authors, limited to max. 10 authors	Production if the author had worked alone
		Authors-per-paper	App	average number of authors per paper over all papers	Collaboration
		Mean pp collaboration	Mean pp collab	Percentage inter-institutional collaboration type, taken from author byline information in Web of Science.	Collaboration
		Mean pp internal collaboration	Mean pp int collab	The proportion of cited references in the publication linking to other WoS publications. A paper with an internal coverage of 0.8%, means that 80% of the references of this paper are covered by the WoS (since 1980)	Cognitivity
Citation-based	c.effect	Percent self-citations	%sc	Number of self-citations divided by total citations	Identifies unwarranted self-promotion
		Percent not cited	%nc	Share of uncited publications	Percentage of work that has not been cited to the present date
	b. normalized	Sum pp top number of citations	Sum pp top n cits	Proportion papers that receive more than 10 citations. 1 is that the paper has more than 10 citations and 0 that is has less	Productivity and impact of a researcher
		Sum pp top prop	Sum pp top prop	Proportion of papers in the top 10% of the world. 100% means that the article belongs to this set of papers, 0 means not.	Identify researcher's papers that are rated top of their field

		Number of productive papers (Antonakis and Lalive, 2008)	NprodP	Used in the Index of Quality and Productivity, a benchmark using the number of years since the researcher defended her doctorate, number of published papers, times cited and top three areas in which the researcher is cited. NprodP is the number of papers that perform better than the benchmark	Amount of papers that are cited more frequently than average papers in the researcher's specialty
		Times cited more than average (Antonakis and Lalive, 2008)	T>ca	Used in the Index of Quality and Productivity and the NprodP indicators: T>ca is the rate the NprodP papers (adjusted papers) perform better than average	How much more than average, as a ratio, the researcher is cited
	a. count	Number of Citations	C	Total number of citations received by publications of the researcher (including self-citations)	Effect of production
		Citations per paper	CPP	The average number of citations per paper, C/P.	Average effect per paper
		Citations minus self-citations	Csc	Total citation count, self-citations removed	Citations from external parties
		Number of self-citations	sc	Sum of self-citations	Building on own research
		Number not cited	nnc	The sum of uncited papers	Non-effectual papers
		Most Significant paper	SIG	The paper with the highest number of citations	Most effectual paper in researcher's portfolio
		Citations less than 5 years old	Cless5	Number of citations less than 5 years old, from the publication of the paper. Publication year is Zero	Currency of citations
		Age Weighted Citation Rate (Harzing, 2012)	AWCR	The number of citations to a given paper divided by the age of that paper. Sum over all papers	Productivity and impact allowing younger, less cited papers to contribute to the index
		Citation age (Egghe and Rousseau, 2000)	Cage	Mean difference between the date of publication of a researcher's work and the age of citations referring to it.	Currency of citations
		Price Index (Price, 1970)	PI	Percentage references to documents, not older than 5 years, at the time of publication of the citing sources	Currency of citations
		Fractional citation count	Fc	Gives an author of an m-authored paper only credit of c/m if the paper received c citations	The effect of each author of a paper
		Fractional Citations per Paper	FracCPP	Fc/Fp	The average effect of each per paper, adjusted for the numbers of author per paper
		Per-author AWCR, (Harzing, 2012)	AWCRpa	AWCR normalized for the number of authors for each paper	The per-author age-weighted citation rate is similar to the plain AWCR, but normalized to the number of authors for each paper.
Journal Impact		Mean citation score	mcs	Mean citation score (journal) self cites not included	Journal impact (prestige of journal the researcher publishes in)

		Mean normalized citation score	mncs	Relates article to world average in regards to document type, publication year and field. 0.9 means cited 10% below average, 1.2% cited 20% above.	Mean normalized citation score (adjusts for field, article type and publication year. SC not included)
		Mean journal score : mean citation score	Mean mjs mcs	Mean citation score of all publishing journals the researcher has published in.	Prestige, benchmark. Expected number of citations of the articles in journals the researchers publish in.
		Maximum journal score mean :citation score	Max mjs mcs	Highest citation score of a journal the researcher has published in	Prestige, most significant place of publication
		Mean normalized journal score	Mean mnjs	Average impact of the journals in which the researcher has published compared to the world citation average in the same subfields	Prestige, corrects for differences among fields
Hybrid	e. time	AR index (Jin et al., 2007)	AR	The square root of the sum of the average number of citations per year of articles included in the h-core, as such the AR index can decrease over time.	Supplement to h index. Accounts for the actual number of citations and age of most productive papers.
		m-quotient, (Hirsch, 2005)	m-quot	h divided by academic age	Productivity and impact of a researcher normalized for academic age of researcher
		mg-quotient	mg-quot	g divided by academic age (Egghe, 2006)	Productivity and impact of a researcher normalized for academic age of researcher
	d. authorship	POP h (Harzing, 2008)	POPh	Divides number of citations by number of authors for that paper, then calculates the h- index of the normalized citation counts	Productivity and impact of a researcher, if the researcher had worked alone.
	c. normalized	Normalized h, (Sidiropoulos et al., 2007)	h-norm	Normalized h=h/np, if h of its np articles have received at least h citations each, and the rest (np-h) articles receive no more than h citations.	Normalizes h-index to compare scientists across fields.
	b. h-independent	Age Weighted h, (Harzing, 2012)	AW	Square root of AWCR, suggested as comparable to the h index	Productivity and impact of researcher, normalized for academic age of researcher
		g-index (g), (Egghe, 2006)	g	Publications are ranked in descending order after number of citations. The cumulative sum of citations is calculated, and where the square root of the cumulative sum is equal to the rank this is g-index	Productivity and impact of a researcher, including highly cited papers
	a. h-dependent	h-index, (Hirsch, 2005)	h	Publications are ranked in descending order after number of citations. Where number of citations and rank is the same, this is the h index	Productivity and impact of a researcher
		$\hbar$, (Millers h) (Miller, 2006)	$\hbar$, Millers_h	Square root of half the total number of citations to all publications	Comparison across field and seniority of papers in the productive core

Q2 (Caberizoa et al., 2012)	Q2	Q2 is the geometric mean of h-index and the median number of citations received by papers in the h-core	Productivity and impact of a researcher. Relates the number of papers to the impact of these papers in the h-core
h2, (Kosmulski, 2006)	h2	Weights most productive papers by finding the cube root of all citations (not just citations to h-core articles).	Productivity and impact of a researcher, including highly cited papers
m index (Bornmann et al., 2008)	m	Median number of citations received by papers in the h-core	Supplement to the h index. Median number of citations to core papers
A index (Jin, 2006; Rousseau, 2006)	A	Average number of citations in h-core thus requires first the determination of h.	Supplement to the h index. Mean number of citations to core papers
e-index, (Zhang, 2009)	e	The e-index is the (square root) of the surplus of citations in the h-set beyond h2, i.e., beyond the theoretical minimum required to obtain a h-index of 'h'.	Supplement to the h index. Production and effect of highly cited papers,
hg (Alonso et al., 2010)	hg	Square-root of (h multiplied by g)	Compare researchers with similar h and g indexes.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.joi.2016.09.003.

References

Äyrämö, S., & Kärkkäinen, T. (2006). *Introduction: To partitioning-based clustering. methods: With a robust example*. Finland: University of Jyväskylä.

Abramo, G., & D'Angelo, C. A. (2011). Evaluating research: From informed peer review to bibliometrics. *Scientometrics*, *87*, 499–514. http://dx.doi.org/10.1007/s11192-011-0352-7

Abramo, G., Cicero, T., & D'Angelo, C. A. (2013). Individual research performance: A proposal for comparing apples to oranges. *Journal of Informetrics*, *7*(2), 528–553.

Abramo, G., D'Angelo, C. A., & Rosati, F. (2014). Career advancement and Scientific performance in universities. *Scientometrics*, *98*, 891–907. http://dx.doi.org/10.1007/s11192-013-1075-8

Aksnes, D. W. (2009). Researchers' perceptions of citations. *Research Policy*, http://dx.doi.org/10.1016/j.respol.2009.02.001

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009). Hg-index: A new index to characterize the scientific output of researchers based on the h- and g-indices. *Scientometrics*, http://dx.doi.org/10.1007/s11192-009-0047-5

Alonso, S., Caberizoa, F. J., Herrera-Viedmac, E., & Herrerac, F. (2010). Hg-index: A new index to characterize the scientific output of researchers based on the h- and g-indices. *Scientometrics*, *82*(2), 391–400.

Antelo, A. A. (2016). SHE Figs 2015. In *Directorate-General for Research and Innovation*.. Resource Document: https://ec.europa.eu/research/swafs/pdf/pub_gender_equality/she_figures_2015-final.pdf Accessed 15.06.16

Antonakis, J., & Lalive, R. (2008). Quantifying scholarly impact: IQP versu the hirsch h. *Journal of the American Society for Information Science and Technology*, *59*(6), 956–969.

Archambault, È., & Gagné, É. V. (2004). *The use of bibliometrics in the Social Science and Humanities Final Report*. Social Sciences and Humanities Research Council of Canada (SSHRCC). Resource document 2004 http://www.science-metrix.com/pdf/SM_2004_008_SSHRC_Bibliometrics_Social_Science.pdf Accessed 15.06.16

Bach, J. F. (2011). *On the proper use of bibliometrics to evaluate individual researchers*. [Resource document 2011 http://www.academie-sciences.fr/activite/rapport/avis170111gb.pdf Accessed 16.05.15

Bacher, J., Pöge, A., & Wenzig, K. (2010). Clusteranalyse: Anwendungsorientierte Einführung. In *Klassifikationsverfahren* (3 ed.). Oldenbourg Verlag.

Bacher, J. (2000). A probabilistic clustering model for variables of mixed type? *Quality & Quantity*, *34*(3), 223–235.

Bloch, C., & Schneider, J. W. (2016). Performance-based funding models and researcher behaviour: An analysis of the influence of the Norwegian Publication Indicator at the individual level. *Research Evaluation*, http://dx.doi.org/10.1093/reseval/rvv047

Bornmann, L., Mutz, R., & Daniel, H. (2008). Are there better indices for evaluation purposes than the h-index? A comparison of nine differnt variants of the h-index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, *59*(5), 830–837.

Chawla, A. (2006). National research priorities in a global perspective: A bibliometric analysis. In *International workshop on webometrics, informetrics and scientometrics & seventh COLLNET meeting*.

Claro, J., & Costa, C. A. V. (2011). A made-to-measure indicator for cross-disciplinary bibliometric ranking of researchers performance? *Scientometrics*, *86*(1), 113–123.

Coe, R. (2002). It's the effect size, stupid: What is effect size and why it is important. *Annual conference of the british educational research association, university of exeter, 12–14 september 2002*,. Availible at: http://www.leeds.ac.uk/educol/documents/00002182.htm Accessed on 23.06.16

Cohen, J. W. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Costas, R., van Leeuwen, T. N., & Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age and productivity and impact. *Journal of the Association for Information Science and Technology*, *61*(8), 1564–1581.

Crespo, J. A., Li, Y., & Ruiz-Castillo, J. (2012). *Differences in citation impact across scientific fields (Working paper economic series 12-06)*. Departamento de Economiá, Universidad Carlos III of Madrid.

Díaz-Faes, A. A., Costas, R., Galindo, M. P., & Bordons, M. (2015). Unravelling the performance of individual scholars: Use of canonical biplot analysis to explore the performance of scientists by academic rank and scientific field. *Journal of Informetrics*, *9*(4), 722–733.
Egghe, L., & Rousseau, R. (2000). Aging, obsolescence, impact, growth and utilization: Definitions and relations. *Journal of the American Society for Information Science and Technology*, *51*(11), 1004–1017.
Egghe, L. (2006). Theories and practise of the g-index? *Scientometrics*, *69*(1), 131–152.
Ellis, P. D. (2009). *Effect size calculators.* , website, availible at: http://www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html Accessed on 23.06.16
Franceschini, F., & Maisano, D. (2011). Criticsm of the hg-index. *Scientometrics*, *86*(2), 339–346.
Frey, B. S., & Rost, K. (2010). Do rankings reflect research quality? *Journal of Applied Economics*, *13*(1), 1–38. http://dx.doi.org/10.1016/S1514-0326(10)60002-5
Glänzel, W., & Moed, H. (2013). Opinion paper: Thoughts and facts on bibliometric indicators? *Scientometrics*, *96*(1), 381–394.
Glänzel, W. (2010). On reliability and robustness of scientometrics indicators based on stochastic models?: An evidence-based opinion paper. *Journal of Informetrics*, *4*(3), 313–319.
Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-Analysis in social research*. London: Sage.
Gower, J. C. (1967). A comparison of some methods: Of cluster analysis. *Biometrics*, *2*(4), 623–628.
Harzing, A.-W., & Alakangas, S. (2016). Google Scholar, scopus and the web of science: A longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*(2), 787–804.
Harzing, H. (2008). *Reflections on the H index. [Electronic version].*. Resource document. http://harzing.com/pop_hindex.html Accessed 15.05.15
Harzing, H. (2012). *Publish or Perish user's manual*.. Resource document. http://www.harzing.com/pophelp/metrics.htm Accessed 15.15
Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 106–128.
Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The leiden manifesto for research metrics. *Nature*, *520*, 429–431. http://dx.doi.org/10.1038/520429a
Hirsch, J. (2005). An index to quantify an indiviual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572.
IBM. (2012). *Cluster evaluation algorithms. In IBM SPSS modeler 15 algorithms guide*. IBM Corporation.
IEEE. (2013). *Appropriate use of bibliometric indicators for the assessment of journals, research proposals, and individuals*.. Resource document http://www.ieee.org/publications_standards/publications/rights/ieee_bibliometric_statement_sept_2013.pdf Accessed 15.15
Ibáñez, A., Larrañaga, P., & Bielza, C. (2013). Cluster methods: For assessing research performance: Exploring Spanish computer science. *Scientometrics*, *97*(3), 571–600.
Iliev, B. Z. (2014). Modern bibliometric indicators and achievements of authors. *Journal of Geometry and Symmetry in Physics*, *33*, 113–128.
Janssens, F., Zhang, L., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-Classification schemes. In J. Ponce, & A. Karahoca (Eds.), *Data mining and knowledge discovery in real life applications* (pp. 89–118). Vienna, Austria: I-Tech.
Jeong, S., & Choi, J. (2012). The taxonomy of research collaboration in science and technology: Evidence from mechanical research through probabilistic clustering analysis. *Scientometrics*, *91*(3), 719–735.
Jin, B. H., Liang, L. L., Rousseau, R., & Egghe, L. (2007). The r and AR indices: Complementing the h-index. *Chinese Science Bulletin*, *52*(6), 855–863.
Jin, B. H. (2006). H-index?: An evaluation indicator proposed by scientist. *Science Focus*, *1*(1), 8–9.
Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis* (1st ed.). New York: Wiley.
Kline, R. B. (2013). Beyond significance testing: Reforming data analysis methods. In *Behavioral research* (2nd ed.). Washington, DC: American Psychological Association.
Kosmulski, M. (2006). A new type Hirsch-index saves time and works equally well as the original h-index? *ISSI Newsletter*, *2*(3), 4–6.
Lancho-Barrantes, B. S. (2010). What lies behind the averages and significance of citation indicators in different disciplines? *Journal of Information Science*, *38*(3), 371–382.
Levitt, J. M., & Thelwall, M. (2014). Is multidisciplinary research more highly cited? A macro-level study. *Journal of the American Society for Information Science and Technology*, *59*(2), 1973–1984.
Liu, X., Yu, S., Moreau, Y., De Moor, B., Glänzel, W., & Janssens, F. (2009). Hybrid clustering of text mining and bibliometrics applied to journal sets. In *Proceedings of the SIAM international conference on data mining, SDM 2009* (pp. 49–60).
MacQueen, J. (1967). Some methods: for classification and analysis of multivariate observations. In L.M. Le Cam, and J. Neyman, (Eds.), Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, (pp. 281–297). Berkeley: University of California Press.
Martin, B. R., & Irvine, J. (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy*, *12*(2), 61–90.
Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, *58*(13), 2105–2125.
Miller, C. W. (2006). *Superiority of the h-index over the impact factor for physics. arXiv.org e-Print archive [Electronic version]*. Available: arXiv:physics/ 2006 0608183.
Otsuki, A., & Kawamura, M. (2013). The study about the analysis of responsiveness pair clustering to social network bipartite graph. *Advanced Computing: An International Journal (ACIJ) AIRCC*, *4*(5), 1–14.
Podlubny, I. (2004). A note on the comparison of scientific impact expressed by number of citations in different fields of science. *Scientometrics*, *64*(1), 95–99.
Price, D. d. S. (1970). Citation measures of hard science, soft science, technology and non-science. In C. E. Nelson, & D. K. Pollack (Eds.), *Communication among scientists and engineers* (pp. 3–22). Lexington: Heath Lexington Books.
Retzer, V., & Jurasinski, G. (2009). Towards objectivity in research evaluation using bibliometric indicators – a protocol for incorporating complexity? *Basic and Applied Ecology*, *10*(5), 393–400.
Rousseau, R. (2006). *New developments related to the Hirsch index e-LIS:e-prints in library and information science*.. Resource document. http://eprints.rclis.org/7616/ Accessed 15.05.15
Ruspini, E. H. (1970). Numerical. methods: For fuzzy clustering. *Information Science*, *2*(3), 319–350.
Sandström, E., Sandström, U., (2009). Meeting the micro-level challenges: bibliometrics at the individual level. In B. Larsen & J. Leta (Eds.), The 12th Conference on Scientometrics and Informetrics, July 14-17, 2009, (pp.856–856), Rio de Janeiro, Brazil.
Schneider, J. W., & Borlund, P. (2007a). Matrix comparison: Part 2: measuring the resemblance between proximity measures or ordination results by use of the Mantel and Procrustes statistics. *Journal of the American Society for Information Science and Technology*, *58*(11), 1596–1609.
Schneider, J. W., & Borlund, P. (2007b). Matrix Comparison: Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal of the American Society for Information Science and Technology*, *58*(11), 1586–1595.
Schneider, J. W., & van Leeuwen, T. D. (2014). Analysing robustness and uncertainty levels of bibliometric performance statistics supporting science policy. A case study evaluating Danish postdoctoral funding. *Research Evaluation*, *23*(4), 285–297. http://dx.doi.org/10.1093/reseval/rvu016
Schneider, J. W. (2013). Caveats for using statistical significance tests in research assessments. *Journal of Informetrics*, *7*(1), 50–62. http://dx.doi.org/10.1016/j.joi.2012.08.005
Sidiropoulos, A., Katsaros, D., & Manopoulos, Y. (2007). Generalized hirsh h-index for disclosing latent facts in citation networks? *Scientometrics*, *72*(2), 253–280.
Sivertsen, G. (2016). Publication-based funding: The norwegian model. In M. Ochsner, S. E. Hug, & H.-D. Daniel (Eds.), *Research assessment in the humanities* (pp. 79–90). Springer International Publishing.
Su, P., Shang, C., & Shen, Q. (2013). Link-based approach for bibliometric journal ranking. *Soft Computing*, *17*(12), 2399–2410.

Sun, X., Tang, W., Ye, T., Zhang, Y., Wen, B., & Zhang, L. (2014). Integrated care: A comprehensive bibliometric analysis and literature review. *International Journal Integrated Care*, 1–12 [e017].
Vinkler, P. (1996). Some practical aspects of the standardization of scientometric indicators? *Scientometrics*, *35*(2), 235–245.
Vinkler, P. (2007). Eminence of scientists in the light of the h-index and other scientometric indicators? *Journal of Information Science*, *33*(4), 481–491.
Waltman, L., Van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks? *Journal of Informetrics*, *4*(4), 629–635.
Wildgaard, L., & Larsen, B. (2014). Scaling analysis of author-level bibliometric indicators. In E. C. M. Noyons (Ed.), *Context counts: Pathways to master big and little data* (pp. 692–701). University of Leiden [3–5 September].
Wildgaard, L., Larsen, B., & Schneider, J. W. (2013). *ACUMEN deliverable 5.3: selection of samples, part 1 & 2*.. Resource document http://iva.ku.dk/ansatte/?pure=da%2Fpersons%2Florna-elizabeth-wildgaard(6e7eb004-dfd6-47b8-aca9-38cb594ceae4)%2Fpublications.html&page= 1 Accessed 15.05.15 May 2015
Wildgaard, L. (2015a). *Measure Up!: The extent author-level bibliometric indicators are appropriate measures of individual researcher performance*. Det Humanistiske Fakultet: Københavns Universitet., 152s. Resourse document: http://iva.ku.dk/ansatte/?pure=da%2Fpublications%2Fmeasure-up(76f0017b-4914-4e9c-8430-8c6b79f01596).html
Wildgaard, L. (2015b). A comparison of 17 author-level bibliometric indicators for researchers in astronomy, environmental science, philosophy and public health in web of science and google scholar. *Scientometrics*, *104*(3), 873–906.
Wouters, P. (2014). *A key challenge: The evaluation gap*.. Resource document https://citationculture.wordpress.com/2014/08/28/a-key-challenge-the-evaluation-gap/ Accessed 15 May 2016
Zhang, C.-T. (2009). The e-index, complementing the h-index for excess citations. *Public Library Of Science*, *4*(5), e5249.
van Arensbergen, P. (2014). *Talent Proof: Selection processes and research funding and careers*. Netherlands: Rathanau Institute.