Journal of Informetrics 9 (2015) 102–117

Contents lists available at ScienceDirect

Journal of Informetrics

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

Field-normalized citation impact indicators using algorithmically constructed classification systems of science

Javier Ruiz-Castillo[a,*], Ludo Waltman[b]

[a] Departamento de Economía, Universidad Carlos III of Madrid, Spain
[b] Centre for Science and Technology Studies, Leiden University, The Netherlands

ARTICLE INFO

Article history:
Received 17 July 2014
Received in revised form 10 October 2014
Accepted 27 November 2014
Available online 18 December 2014

Keywords:
Field normalization
Classification systems
Clustering methodology
Citation impact indicators
University rankings

ABSTRACT

We study the problem of normalizing citation impact indicators for differences in citation practices across scientific fields. Normalization of citation impact indicators is usually done based on a field classification system. In practice, the Web of Science journal subject categories are often used for this purpose. However, many of these subject categories have a quite broad scope and are not sufficiently homogeneous in terms of citation practices. As an alternative, we propose to work with algorithmically constructed classification systems. We construct these classification systems by performing a large-scale clustering of publications based on their citation relations. In our analysis, 12 classification systems are constructed, each at a different granularity level. The number of fields in these systems ranges from 390 to 73,205 in granularity levels 1–12. This contrasts with the 236 subject categories in the WoS classification system. Based on an investigation of some key characteristics of the 12 classification systems, we argue that working with a few thousand fields may be an optimal choice. We then study the effect of the choice of a classification system on the citation impact of the 500 universities included in the 2013 edition of the CWTS Leiden Ranking. We consider both the MNCS and the $PP_{top\ 10\%}$ indicator. Globally, for all the universities taken together citation impact indicators generally turn out to be relatively insensitive to the choice of a classification system. Nevertheless, for individual universities, we sometimes observe substantial differences between indicators normalized based on the journal subject categories and indicators normalized based on an appropriately chosen algorithmically constructed classification system.

1. Introduction

In this paper, we deal with the problem of normalizing citation impact indicators based on a classification system of science. As we know, the choice of a *classification system*, that is, the assignment of individual scientific publications (or journals) to research areas, remains an open question in Scientometrics. Together with the well-known classification systems included in Thomson Reuters' Web of Science (WoS hereafter) and Elsevier's Scopus databases, there are a number of interesting proposals suggested by individual researchers (see inter alia the references in Waltman & Van Eck, 2012).

In practice, the choice of the WoS classification system is often made because it is the only classification system that is readily available. However, a number of studies question the appropriateness of the WoS classification system for the purpose

* Corresponding author. Tel.: +34 91 624 95 88.
E-mail address: jrc@eco.uc3m.es (J. Ruiz-Castillo).

http://dx.doi.org/10.1016/j.joi.2014.11.010
1751-1577/

of normalizing citation impact indicators. Neuhaus and Daniel (2009) contrast the assignment of individual publications to WoS subject categories based on the journal where they have appeared with a novel methodology for Chemistry and related fields where each publication is directly assigned to one of the 80 sections of the *Chemical Abstracts* database. Taking the journal *Angewandte Chemie* as an example, they illustrate the limitations of the WoS journal classification scheme in the case of general journals. On the other hand, using the 20 sections under the Biochemistry heading they clearly illustrate that citation habits vary extensively not only *between* fields but also *within* fields. Similarly, Van Eck, Waltman, Van Raan, Klautz, and Peul (2013) establish the existence of heterogeneous sub-groups (corresponding to clinical and basic medical research) with different citation practices within WoS subject categories. In this case, the mean citation of an entire category is simply the weighted average of different, and hence non-comparable sub-group mean citations. This is exactly the same problem found by Van Leeuwen and Calero-Medina (2012) inside the Thomson Reuters broad field of Economics and Business. Within the dominant WoS journal subject category in that field, denoted *Economics*, these authors find strong differences across 19 specialties defined in the EconLit electronic bibliography produced by the American Economic Association. Finally, Leydesdorff and Bornmann (2014) point out that the WoS subject categories were developed decades ago for the purpose of information retrieval and evolved incrementally with the database; the classification is machine-based and partially manually corrected. This contribution shows the potential problems for research evaluation in one discipline that is attributed a WoS category – *Information Science and Library Science* – and one specialty which is not – *Science and Technology Studies*.

Clearly, the comparison of the WoS system with some relevant alternatives is an important research problem. In this paper, we search for alternatives within the publication-level algorithmic methodology introduced by Waltman and Van Eck (2012) (the WVE methodology hereafter). This methodology is able to handle very large datasets, and uses a transparent clustering technique that classifies publications into clusters solely based on direct citations between them. Contrary to the WoS system, each publication is assigned to a single cluster. Moreover, the WVE methodology can be used to construct classification systems that, unlike for instance the *Chemical Abstracts* and EconLit systems, cover all scientific fields.

In the first large-scale application of the WVE methodology (Waltman & Van Eck, 2012), three types of parameter values needed to be chosen: the number of what we call *granularity levels* and, at each level, the *minimum number of publications per cluster*, and the *resolution parameter* that determines the level of detail of the clustering (i.e., a small number of large clusters vs. a large number of small clusters). In this paper, we consider a set of twelve granularity levels that are not restricted to be hierarchically linked. Thus, by fixing the resolution parameter at twelve different values, we build a sequence of independent classification systems in each of which the same set of publications is assigned to an increasing number of clusters. Furthermore, no minimum number of publications per cluster is imposed at any granularity level. Thus, at every step, the WVE algorithm freely determines a cluster size distribution.

We apply this scheme to a WoS dataset consisting of 3.6 million articles published in 2005–2008 in academic journals – excluding trade journals, national journals, etc. – and the citations they receive during a five-year citation window for each year in that period. The number of clusters in the WVE sequence ranges from 390 to 73,205 in granularity levels 1–12. This contrasts with the 236 clusters (i.e., journal subject categories) in the WoS classification system.

Which granularity level is used in practice in the calculation of normalized citation impact indicators is a very important issue. As clearly argued by Zitt, Ramana-Rahari, and Bassecoulard (2005), "*An article may exhibit very different citation scores or rankings when compared within a narrow specialty or a large academic discipline.*"(op. cit., p. 391). If we choose a granularity level dominated by a relatively small number of broad fields, the danger is that they are too heterogeneous, in which case comparisons between publications within the same cluster may be biased. For instance, this may affect the *Essential Science Indicators* of Thomson Reuters that provide reference standards solely for 22 broad fields of research. However, when we go in the opposite direction and choose a classification system including too many clusters, we face difficulties of a different nature. Firstly, some clusters may mostly include the output of a subset of closely connected authors citing each other, and isolated from *bona fide* scientific communities whose output is classified in other clusters. Secondly, some clusters may be so small as to jeopardize their statistical properties. Thirdly, some clusters may have artificially low mean citations, so that standard normalization procedures that use cluster mean citations as normalization factors will tend to over-estimate these clusters' publications against those in high impact clusters characterized by a high mean citation. It may very well be the case that classification systems characterized by high granularity levels are plagued with clusters that present the above three difficulties together.

As a consequence of the above issues, the evaluation of research units based on citation impact is likely to be dependent on the granularity level at which the evaluation takes place. As Zitt et al. (2005) conclude, "*The fact that citation indicators are not stable from a cross-scale perspective is a serious worry for bibliometric benchmarking. What can appear technically as a 'lack of robustness' raises deeper questions about the legitimacy of particular scales of observation.*" (op. cit., p. 392). Adams, Gurney, and Jackson (2008) reach a similar conclusion: "*the fact that more than one view and hence more than one interpretation of performance might exist would need to be taken into account in any evaluation methodology*" (op. cit., p. 94). For other studies on this problem, we refer to Colliander and Ahlgren (2011) and Glänzel, Thijs, Schubert, and Debackere (2009).

In this paper, we investigate two questions. Firstly, what are the main characteristics of the twelve WVE classification systems, and how do they compare with those of the WoS alternative? Secondly, what are the consequences of using the WoS classification system or an appropriately selected member of the WVE sequence for the evaluation of the citation impact of universities?

For the first purpose, we study how the following characteristics evolve as the granularity level increases: the cluster size and the cluster mean citation distributions, the degree of skewness and the similarity of this characteristic across cluster

citation distributions, and the degree of homogeneity within cluster citation distributions. For the second purpose, we analyze the more than 1.8 million articles – about 50% of the total – corresponding to the 500 universities in the 2013 edition of the CWTS Leiden Ranking (Waltman, Calero-Medina, et al., 2012). We use a fractional approach to solve the classical assignment problem of individual publications to several WoS categories. This problem does not affect algorithmically constructed classification systems, since in these systems each publication is assigned to a single cluster. We also use a fractional counting approach to solve the problem – present in all classification systems – of publications assigned to several co-authors working in different institutions. We follow two evaluation criteria. Firstly, we use the Mean Normalized Citation Score indicator (MNCS hereafter, Waltman, Van Eck, Van Leeuwen, Visser, & Van Raan, 2011), where normalization is performed at the cluster level in each classification system. Secondly, we believe that it is important to extend the analysis to the members of the percentile rank approach (see Bornmann & Marx, 2013, for a summary of this approach and some of the recent literature related to it). In particular, we use the $PP_{top\ 10\%}$ indicator, defined as the percentage of an institution's scientific output included in the set formed by the 10% of the most highly cited publications in their respective scientific fields. This indicator is included in the influential Leiden and *SCImago* rankings.[1]

The remainder of the paper is organized into four Sections. Section 2 discusses some characteristics of the WoS and WVE classification systems, while Sections 3 and 4 present the results on the citation impact of 500 universities under different classification systems using the MNCS and the $PP_{top\ 10\%}$ indicator, respectively. Section 5 summarizes the paper, discusses the main findings, and suggests some extensions. We note that a more extensive version of this paper is available as a working paper (Ruiz-Castillo & Waltman, 2014; RCW hereafter). Below, we will sometimes refer to this working paper for additional results that, for reasons of space limitations, are not reported in full detail in the present paper.

2. Characteristics of the different classification systems

2.1. Building the twelve WVE classification systems

In the first large-scale application of their approach, Waltman and Van Eck (2012) classify almost ten million documents (of the type article, letter, and review) published in the sciences and social sciences in the WoS database during the period 2001–2010. The choice of parameter values determines a classification system that distinguishes between three granularity levels with a minimum of 120,000, 5000, and 50 publications per cluster, respectively. The three levels are nested, or hierarchically ordered in the sense that the 22,412 clusters in level 3 are a partition of and induce the 672 clusters in level 2, whereas the clusters in level 2 in turn are a partition of and induce the 20 clusters in level 1. As pointed out in the Introduction, in this paper we apply the WVE methodology by merely changing the resolution parameter that essentially determines the number of clusters at each granularity level. Requiring clusters to be nested across granularity levels would restrict the working of the algorithm by imposing dependencies between granularity levels. Therefore, by not imposing a hierarchical structure we achieve what we want, namely, to focus exclusively on the consequences of changing the granularity level. Similarly, fixing the minimum number of articles per cluster would restrict the way the algorithm performs at each granularity level. Thus, not imposing any restriction in this direction allows the algorithm to freely determine the cluster size distribution.[2]

The data source for the analyses is the WoS database produced by Thomson Reuters. Since we wish to address a homogeneous population, only publications of the WoS document types article and review are considered. In the rest of this paper, we refer to publications of these two document types simply as 'articles' or 'publications'. However, following Waltman and Van Eck (2013a, 2013b), we exclude publications in local journals, as well as popular magazines and trade journals. We work with journals in the sciences, the social sciences, and the arts and humanities, although many arts and humanities journals are excluded because they are of a local nature. We apply the WVE methodology twelve times to 9,446,622 publications from the period 2003–2012, but we then truncate each cluster to include only publications from the period 2005–2008. The reason for first applying the WVE methodology to a ten-year time period and then truncating the clusters to include only four years of publications is that the WVE methodology relies on direct citations rather than bibliographic coupling or co-citations and therefore requires a sufficiently long time period in order to produce high-quality clusters. The number of publications in the period 2005–2008 is 3,614,447.

Two points should be made concerning our application of the WVE methodology. Firstly, as part of this methodology, a large-scale optimization problem needs to be solved. For this purpose, we use the so-called smart local moving algorithm, recently proposed by Waltman and Van Eck (2013c), and freely available at www.ludowaltman.nl/slm/. Secondly, in the original WVE methodology there are some publications that cannot be assigned to a cluster because they have no citation relations with other publications (Waltman & Van Eck, 2012). In the methodology that we use in this paper, publications in this situation are assigned to a cluster based on the journal in which they have appeared. More precisely, publications are

[1] SCImago is a research group from the *Consejo Superior de Investigaciones Científicas*, University of Granada, Extremadura, Carlos III (Madrid) and Alcalá de Henares in Spain. The *SCImago Institutions Rankings* (SIR; www.scimagoir.com) is a bibliometric ranking of research institutions based on Elsevier's Scopus database.

[2] In practical applications of our work, one may need to impose a minimum cluster size condition. This can be done following the approach described by Waltman and Van Eck (2012).

assigned to the cluster in which their journal has most publications. In this way, publications without citation relations can still be assigned to a cluster and there are no publications without a cluster assignment.

The rest of this section addresses the following questions:

(i) How do cluster size and cluster mean citations evolve as the granularity level increases in the WVE sequence? How do the distributions of these variables in WVE classification systems compare with the corresponding distributions in the WoS system?
(ii) Do cluster citation distributions comply with the stylized features established in previous research? That is, are cluster citation distributions highly skewed and, in addition, are they very similar across clusters in all classification systems?
(iii) The conjecture is that the greater the number of clusters, the more homogeneous cluster citation distributions will become. Is there any evidence concerning this trend in our classification systems?

Ideally, the answers to these questions will help us to select a number of admissible granularity levels in the WVE sequence with which to take on the ranking of universities using the WoS system.

2.2. The joint variation of cluster size and cluster mean citation in the different classification systems

For any classification system, we sort clusters in decreasing order by size, where size is measured as the number of publications, and group clusters into ten decile classes. For each decile, the results concerning the average number of publications per cluster, denoted by μ, and the average number of citations per publication, denoted by MC, are in Table 1. The number of citations of a publication is counted during a five-year citation window. For instance, in the case of a publication from 2006, citations are counted during the period 2006–2010. Together with the information for the WoS system and the twelve members of the WVE sequence, for reference purposes we have included in Table 1 a granularity level 13 where each article forms an independent cluster.

Let us first analyze the sequence of WVE classification systems. In the nine deciles 2–10 in level 1, μ ranges from eleven to one single publication per cluster, while MC is equal to only one or less citations per article. Next, consider an intermediate classification system such as level 6. The above characteristics are now present only in deciles 8–10. Finally, as expected, when the resolution parameter reaches the highest value in level 12, μ becomes less than 100 publications in all deciles but the first. However, mean citation per article ranges from 5.0 to 11.7 citations. What do we observe in the WoS system? All deciles consist of relatively large clusters and, except for the last three, MC is greater than five citations per article.

We emphasize the following three conclusions. Firstly, as the resolution parameter increases, the number of deciles in which μ is greater than 1000 in Table 1 follows an inverted U trend: it ranges from one in levels 1 and 2, up to five in level 6. From this level until the end of the WVE sequence, the number of deciles with "large" clusters diminishes to one in level 9, and none in levels 10–12. Secondly, the main difference between the WVE sequence and the WoS system is the presence in the former of a large number of small clusters (typically accompanied by a low mean citation per article). If we define *small* as less than or equal to 100 publications per cluster, the number of such clusters in the WoS system is only five, whereas in the WVE sequence it ranges from a few hundred in levels 1–7 up to 64,375 in level 12 (see row B in Table 1).[3] Thirdly, it is important to note that, up to level 8 in the WVE sequence, the set of small clusters includes a very small proportion (i.e., less than 0.9%) of the 3.6 million articles in the entire dataset (see row D in Table 1). From levels 9 to 12, however, this percentage increases from 3.2% to 61.3% of the total.

This analysis suggests that the use of granularity levels 9–12 in the calculation of normalized citation impact indicators may be problematic. At the same time, these results lead us to investigate the situation when we restrict the attention to *significant* clusters with at least 100 publications (see row C in Table 1). For levels 1–8, the number of significant clusters increases monotonically from 17 to 4161 clusters. In the WoS system this number is 231, very close to the 228 clusters in level 4. For reasons of space, results similar to the ones reported in Table 1 but for significant clusters only can be found in Table 1B in RCW. The main conclusions based on this table can be summarized as follows. Firstly, granularity levels 9–12 are still dominated by relatively small clusters. In comparison, the cluster distribution for levels 1–8 is more appealing. Secondly, the cluster size distribution in level 4 is now comparable to the one in the WoS system. This illustrates that the main difference between the WoS system and level 4 is that the way journals are assigned to subject categories in the former neglects to recognize a key feature of science: the presence, in different degrees, of small clusters, or what we may call *small science*. Thirdly, for lower granularity levels, there is a large percentage of clusters with a number of publications per cluster greater than this quantity for the first decile in the WoS system, while the opposite is the case for granularity levels above level 4: there is a large percentage of clusters in the WoS system with a number of publications per cluster greater than this quantity in the first decile in the relevant WVE systems.

[3] The number of small clusters does not increase monotonically as the granularity level increases. In particular, the relatively small number of small clusters at granularity level 3 is somewhat remarkable. This is probably due to issues with local optima in the optimization problem that is solved in the WVE methodology.

Table 1
Mean number of publications per cluster, μ, and mean citation per publication, MC, in the partition by deciles of the cluster distribution for the WoS system and the twelve WVE granularity levels.

	WoS system		Level 1		Level 2		Level 3		Level 4	
	μ	MC	μ	MC	μ	MC	μ	MC	μ	MC
Deciles										
1	58,892.5	9.7	92,643.6	8.7	73,730.7	8.7	75,780.6	9.3	34,806.9	9.3
2	31,493.8	10.5	10.9	1.0	11.8	1.0	25,446.7	7.1	15,901.2	8.4
3	20,298.4	8.2	6.4	0.6	5.6	0.5	2,820.4	6.3	6,569.5	6.8
4	13,840.0	6.2	4.5	0.5	4.1	0.5	13.0	0.8	1,007.7	4.5
5	10,099.8	6.3	3.4	0.4	3.1	0.6	5.7	0.4	13.3	1.0
6	6,915.9	6.0	3.0	0.4	3.0	0.4	4.1	0.6	5.8	0.5
7	4,454.8	5.5	2.2	0.7	2.0	0.7	3.0	0.4	3.9	0.3
8	2,849.3	5.0	2.0	0.4	1.9	0.5	2.3	0.6	3.0	0.3
9	1,663.4	4.0	1.0	0.7	1.0	0.6	1.6	0.4	2.0	0.5
10	488.0	4.4	1.0	0.4	1.0	0.5	1.0	0.5	1.0	0.6
A. No. of clusters	236		390		489		341		613	
B. No. of small clusters	5		373		450		248		385	
C. Number of significant clusters	231		17		39		93		228	
D. % of publications in small clusters			0.05%		0.06%		0.04%		0.08%	

	Level 5		Level 6		Level 7		Level 8		Level 9	
	μ	MC	μ	MC	μ	MC	μ	MC	μ	MC
Deciles										
1	19,267.0	9.4	10,432.5	9.9	4,584.6	9.8	2,472.7	10.5	1,326.5	10.9
2	9,241.8	8.9	5,977.2	8.3	2,714.2	9.3	1,435.3	9.4	758.8	9.3
3	5,711.5	7.2	4,089.0	8.4	1,934.9	8.8	1,015.9	8.0	527.8	8.1
4	2,841.3	6.7	2,795.8	7.8	1,353.4	7.6	737.3	7.4	385.9	7.3
5	655.7	5.3	1,809.6	7.0	965.8	6.7	542.1	6.9	284.4	6.3
6	17.5	1.2	976.1	6.1	661.2	6.2	377.4	5.8	205.8	5.6
7	5.3	0.6	321.8	4.8	406.8	5.3	250.8	5.1	143.9	5.0
8	3.3	0.5	19.2	1.5	201.1	4.4	151.3	4.5	96.2	4.2
9	2.1	0.4	3.4	0.5	43.1	3.0	70.5	3.7	57.7	3.7
10	1.1	0.4	1.4	0.5	2.4	0.5	6.0	1.2	14.0	3.0
A. No. of clusters	952		1,363		2,805		5,119		9,506	
B. No. of small clusters	482		411		533		958		2,440	
C. Number of significant clusters	470		952		2,272		4,161		7,066	
D. % of publications in small clusters	0.09%		0.10%		0.27%		0.89%		3.2%	

	Level 10		Level 11		Level 12		Level 13	
	μ	MC	μ	MC	μ	MC	μ	MC
Deciles								
1	580.5	11.3	313.2	11.5	169.4	11.7	1.0	9.4
2	317.2	9.4	167.7	9.4	88.8	9.3	1.0	8.0
3	222.7	7.9	116.3	7.9	62.1	7.9	1.0	8.7
4	163.4	7.0	85.3	6.8	46.3	6.6	1.0	8.3
5	120.4	6.1	64.3	6.0	35.8	5.9	1.0	8.1
6	88.7	5.2	49.4	5.1	28.5	5.4	1.0	9.4
7	65.9	4.7	38.1	4.6	23.0	5.0	1.0	8.3
8	48.3	4.1	29.5	4.4	18.3	5.0	1.0	9.5
9	33.5	4.0	21.8	4.3	13.9	4.9	1.0	9.0
10	13.6	3.5	11.0	4.1	7.5	5.0	1.0	7.8
A. No. of clusters	21,849		40,305		73,205		3,614,447	
B. No. of small clusters	10,677		28,318		64,375			
C. Number of significant clusters	11,172		11,987		8,830			
D. % of publications in small clusters	14.4%		33.7%		61.3%			

2.3. *The skewness of science in the different classification systems*

As originally suggested in Price (1965) and afterwards analyzed in Seglen's (1992) seminal contribution, it has been known for some time that citation distributions in different contexts are highly skewed. In addition, using large WoS datasets of field citation distributions at different levels of aggregation and different citation window lengths, recent research has provided convincing evidence concerning two fundamental facts: citation distributions are not only highly skewed but, very importantly, very similar across scientific fields (Albarrán & Ruiz-Castillo, 2011; Albarrán et al., 2011; Glänzel, 2007; Li, Castellano, Radicchi, & Ruiz-Castillo, 2013; Radicchi & Castellano, 2012; Radicchi, Fortunato, & Castellano, 2008; Waltman,

Table 2
The skewness of cluster citation distributions according to the CSS approach. Average (standard deviation), and coefficient of variation (CV) over significant clusters at all granularity levels of the percentages of articles, and the percentages of total citations by category. All classification systems. A cluster is considered significant if it has at least 100 publications.

Classification system	Number of significant clusters	Percentage of articles in category			Percentage of citations in category		
		1	2	3	1	2	3
WoS	231	69.0 (3.3)	21.5 (2.0)	9.5 (1.7)	23.0 (3.9)	33.5 (1.8)	43.4 (3.8)
CV		0.05	0.09	0.18	0.17	0.05	0.09
0	1	72.0 (0.0)	20.2 (0.0)	7.8 (0.0)	22.6 (0.0)	32.3 (0.0)	45.2 (0.0)
1	17	70.9 (3.1)	20.7 (2.0)	8.4 (1.3)	22.8 (3.7)	33.0 (1.7)	44.2 (3.4)
CV		0.04	0.10	0.16	0.16	0.05	0.08
2	39	70.3 (3.6)	21.2 (2.6)	8.5 (1.3)	(5.1)	33.7 (2.5)	44.5 (3.6)
CV		0.05	0.12	0.15	0.24	0.07	0.08
3	93	70.3 (3.2)	21.1 (2.0)	8.6 (1.5)	22.4 (3.6)	33.7 (2.2)	43.9 (3.3)
CV		0.05	0.09	0.17	0.16	0.07	0.07
4	228	70.3 (3.3)	21.0 (2.0)	8.7 (1.7)	22.7 (4.5)	33.4 (2.0)	43.9 (3.9)
CV		0.05	0.09	0.19	0.20	0.06	0.09
5	470	69.7 (3.4)	21.3 (2.1)	9.1 (1.7)	23.0 (4.1)	33.4 (2.1)	43.5 (3.8)
CV		0.05	0.10	0.19	0.18	0.06	0.09
6	952	69.4 (3.7)	21.2 (2.2)	9.3 (1.9)	23.3 (4.0)	33.4 (2.3)	43.3 (3.9)
CV		0.05	0.11	0.20	0.17	0.07	0.09
7	2272	68.7 (4.0)	21.5 (2.5)	9.8 (2.1)	23.2 (4.5)	33.4 (2.7)	43.4 (4.3)
CV		0.06	0.11	0.22	0.19	0.08	0.10
8	4161	68.3 (4.2)	21.7 (2.7)	10.0 (2.3)	23.3 (4.6)	33.5 (3.0)	43.1 (4.5)
CV		0.06	0.12	0.23	0.19	0.09	0.10
9	7066	67.8 (4.4)	21.8 (2.8)	10.4 (2.5)	23.5 (4.7)	33.5 (3.2)	43.0 (4.8)
CV		0.07	0.13	0.24	0.20	0.10	0.11
10	11,172	67.3 (4.6)	22.1 (3.0)	10.6 (2.7)	24.1 (4.9)	33.4 (3.5)	42.5 (5.0)
CV		0.07	0.14	0.25	0.20	0.11	0.12
11	11,987	67.1 (4.8)	22.2 (3.2)	10.7 (2.9)	24.6 (5.1)	33.3 (3.8)	42.1 (5.4)
CV		0.07	0.14	0.27	0.21	0.11	0.13
12	8830	67.2 (5.2)	22.1 (3.4)	10.7 (3.1)	24.9 (5.9)	33.2 (4.0)	42.0 (6.0)
CV		0.08	0.15	0.28	0.24	0.12	0.14

Calero-Medina, et al., 2012; Waltman, Van Eck, et al., 2012).[4] This similarity has opened up the way for the justification of meaningful comparisons of citation impact across heterogeneous fields (Crespo, Li, & Ruiz-Castillo, 2013; Crespo, Herranz, Li, & Ruiz-Castillo, 2014; Li et al., 2013; Glänzel, 2011; Radicchi & Castellano, 2012; Radicchi et al., 2008; Ruiz-Castillo, 2014).

In this context, our next question is whether cluster citation distributions for all classification systems follow the same pattern that has been found in this literature. One convenient way to approach this issue is to apply the Characteristic Scales and Scores (CSS hereafter) size- and scale-independent technique, first used in Scientometrics in Schubert, Glänzel, and Braun (1987), and also used in some of the above references. For that purpose, the following two *characteristic scores* are determined for every cluster in every classification system: m_1 = mean citation of the cluster citation distribution, and m_2 = mean citation for articles with a number of citations above m_1. Consider the partition of any cluster distribution into three broad classes: (i) articles with low impact, or a number of citations less than or equal to m_1; (ii) articles with a fair impact, or citations greater than m_1 and less than or equal to m_2; (iii) articles with a remarkable or outstanding citation impact above m_2. For each significant cluster (with at least 100 publications), we compute the percentage of articles in the three classes, and the corresponding percentages of the total number of citations accounted for by each class. The average (the standard deviation), and the coefficient of variation of the six values over all significant clusters for every classification system appear in Table 2. For reference, we have included the case in which all articles are included in a single cluster, say the overall citation distribution for the entire dataset, as if the citations received by the 3.6 million articles were comparable. This case is denoted as level 0.

The results are remarkable in several respects. Firstly, the average percentages of articles in each class – approximately equal to 69–70%/21%/9–10% – illustrate the high skewness of cluster citation distributions, while the relatively low standard deviations and coefficients of variation show the strong similarity across clusters. These two features – high skewness and strong similarity of cluster citation distributions – are typically found in the literature on citation distributions using large WoS datasets.[5] Secondly, the skewness of science is already present when all articles are grouped in a single cluster in level 0. Therefore, for this phenomenon to appear it is not necessary that articles are appropriately classified into conventional scientific fields at different levels of aggregation. Thirdly, cluster citation distributions for levels 1–6 present essentially the

[4] In the same vein, for the skewness and similarity of individual scientists' productivity distributions across 30 broad fields, see Ruiz-Castillo and Costas (2014).

[5] For example, for 3.7 million articles published in 1998–2002 in 219 WoS subject categories with a five-year citation window, 68.6% of articles are poorly cited, so that they account for only 29.1% of all citations, while 10% of very highly cited articles account for 44.9% of all citations (Table 1 in Albarrán et al., 2011).

Table 3
Between-group citation inequality as a percentage of overall citation inequality.

Classification system	Between-group citation inequality, % overall citation inequality
WoS	15.9
1	6.8
2	8.8
3	9.7
4	11.3
5	12.8
6	15.1
7	18.8
8	20.9
9	23.8
10	27.8
11	31.1
12	34.7

same average pattern, including relatively small standard deviations and coefficients of variation. For granularity levels 7–12, average cluster skewness is slightly smaller. Moreover, as average cluster size in all deciles becomes smaller (see Table 1) and within-cluster variability increases, we observe that the similarity across clusters is somewhat less striking. Recall that the WVE algorithm classifies publications into clusters on the basis of direct citations between them. We find it reassuring that, under this single restriction, significant clusters at every granularity level reproduce in an acceptable way the skewness of science already documented in WoS classification systems.

What is the situation when we include small clusters in the exercise? To save space, results for selected granularity levels when clusters have at least ten articles are in Table 2B in RCW. On average, the skewness is still intact. However, standard deviations and coefficients of variation are much larger than before. Therefore, we conclude that the WVE algorithm is able to capture well the similarity across cluster citation distributions when we restrict the analysis to significant clusters with at least 100 publications. Recall that for levels 1–8 this means setting apart less than 0.9% of the 3.6 million articles in the entire dataset – clearly, a tolerable restriction.

2.4. Cluster homogeneity in the different classification systems

As indicated in Section 1, we are concerned about the possible lack of comparability, or lack of homogeneity between articles in any cluster in any classification system. In Van Eck et al. (2013), the authors had a priori information about this possibility in a number of subject categories within the WoS system. The problem, of course, is that we do not have any information about which clusters may lack the desirable homogeneity within any given classification system. Nevertheless, as explained in detail in the Appendix in RCW, under the reasonable assumption that as the granularity level and the number of clusters increase, the degree of homogeneity also increases, we can use an additively decomposable citation inequality index to approximate the degree of homogeneity at every granularity level.

The main idea developed in the Appendix in RCW is as follows. Let $\boldsymbol{C}$ be the original citation distribution, and let I be the Theil citation inequality index. Given a partition of $\boldsymbol{C}$ into the set of clusters at a granularity level g, the overall citation inequality $I(\boldsymbol{C})$ can be decomposed into two terms, one capturing the within-group citation inequality, ${I_g}^W$, and another capturing the between-group citation inequality, ${I_g}^B$. Under the assumption that cluster homogeneity tends to increase as the granularity level increases, we must have that ${I_g}^B$ is smaller than ${I_{g+1}}^B$ at the next granularity level $(g+1)$. Correspondingly, ${I_g}^W$ should be greater than ${I_{g+1}}^W$. Therefore, the ratio ${I_g}^B/I(\boldsymbol{C})$ can be taken as a measure of the degree of homogeneity at granularity level g. The value of this ratio for every classification system is in Table 3.

The following two comments should be made. Firstly, within the WVE sequence we confirm that, relative to the overall citation inequality for the entire dataset, the percentage represented by between-group citation inequality increases with the granularity level. However, we must recognize that our measure of homogeneity increases in nearly constant steps as we move from granularity level 0 to level 13. In other words, no granularity level seems to be particularly privileged on this account. Secondly, the degree of homogeneity improvement associated with the move from level 0 to the WoS system is quite large. The conclusion from this analysis is that, within the WVE sequence, we should focus on granularity levels equal to or greater than level 6 to achieve at least a comparable degree of homogeneity as the WoS system itself.

2.5. Recommended granularity levels within the WVE sequence

Is it possible to select an optimal granularity level from the WVE sequence as an alternative to the WoS system for the calculation of normalized citation impact indicators? Although the results presented so far do not provide us with a clear criterion to single out an optimal granularity level, we can give some arguments for recommending certain options in that sequence.

Firstly, what we refer to as the smallness problem leads us to reject levels 9–12 in favor of the others. Secondly, levels 1–6 seem to perform slightly better than levels 7–12 in capturing the skewness of science. However, large clusters that are

Table 4
The skewness of university MNCS distributions according to the CSS approach, and the coefficient of variation of university MNCS values.

	M_1	M_2	Proportion of universities in category			Coefficient of variation
			1	2	3	
WoS	1.02	1.23	0.54	0.30	0.17	0.27
0	1.01	1.30	0.53	0.29	0.18	0.35
1	1.02	1.27	0.56	0.28	0.16	0.29
2	1.02	1.26	0.53	0.30	0.17	0.29
3	1.02	1.25	0.52	0.30	0.18	0.28
4	1.01	1.24	0.52	0.30	0.18	0.28
5	1.01	1.23	0.51	0.30	0.19	0.27
6	1.01	1.22	0.50	0.32	0.18	0.27
7	1.01	1.21	0.50	0.30	0.19	0.25
8	1.01	1.19	0.49	0.31	0.20	0.23
9	1.01	1.18	0.50	0.31	0.19	0.22
10	1.01	1.17	0.49	0.31	0.20	0.20
11	1.01	1.16	0.50	0.31	0.19	0.19
12	1.01	1.15	0.49	0.31	0.20	0.18
13	1.00	0.00	1.00	0.00	0.00	0.00

M_1 = mean university MNCS value.
M_2 = mean MNCS value for universities with MNCS value greater than M_1.
Category 1 = universities with a low MNCS value, less than or equal to M_1.
Category 2 = universities with a fair MNCS value, greater than M_1 and less than or equal to M_2.
Category 3 = universities with a remarkable or outstanding MNCS value, greater than M_2.

bound to be too heterogeneous overtly dominate the former levels. As a matter of fact, it is only from level 6 onwards that we get a measure of homogeneity similar to or better than the one associated to the WoS system. In conclusion, for the purpose of normalizing citation impact indicators in the next two sections, we believe that it is sensible to use levels 7 and 8 within the WVE sequence. At these levels, we have 2805 and 5119 clusters, of which 2272 and 4161 are significant clusters with at least 100 publications. In brief, in levels 7 and 8 most clusters are significant, the percentage of articles in small clusters is smaller than 1%, they clearly show a greater homogeneity than the WoS system for which they are supposed to provide an alternative, and they still capture in an acceptable way the skewness of science across clusters.

3. The citation impact of universities under different classification systems according to the MNCS indicator

This section has two aims. The first aim is to explore some general features of the variation of the citation impact of universities according to the MNCS indicator (Waltman et al., 2011) under different classification systems.[6] The second aim is to perform a direct comparison of the citation impact of universities for the WoS system and the two WVE granularity levels selected for this purpose in Section 2. The analysis reported in this section is based on the 500 universities included in the 2013 edition of the CWTS Leiden Ranking. The information on universities' MNCS values for all classification systems, as well as the country and the number of publications of each university can be found in Table A in the Appendix in RCW, where universities are ordered according to the MNCS results for granularity level 8.

3.1. Variation under different classification systems

Some summary measures for university MNCS distributions can be found in Table 4. Like in Section 2, we use the CSS approach. Column 1 includes the average of university MNCS values, M_1, for each classification system, while column 2 includes the average of MNCS values for universities with an MNCS value greater than M_1, denoted M_2. In order to assess the skewness of each university MNCS distribution, in columns 3–5 we partition the distribution into three classes for universities with an MNCS value smaller than or equal to M_1, greater than M_1 and smaller than or equal to M_2, and greater than M_2, respectively. Finally, column 6 includes the coefficient of variation of university MNCS values, which offers a measure of university MNCS inequality.

[6] We note that in the calculation of the university MNCS values we have normalized only for field, not for publication year. This is different from the way in which MNCS calculations are performed in the CWTS Leiden Ranking. However, since in this paper we work with a fixed-length citation window instead of a variable-length one, normalization for publication year may be considered less important. We also note that in the assignment of publications to institutions a fractional counting approach is adopted. Hence, publications co-authored by multiple institutions are assigned fractionally to each institution. An argument in favor of a fractional rather than a full counting approach is provided by Waltman, Calero-Medina, et al. (2012). Finally, it should be mentioned that in the algorithmically constructed classification systems all clusters, both the significant and the small ones, are included (results restricted to significant clusters only, which are essentially indistinguishable, are available on request). An alternative approach would be to get rid of the small clusters by merging them with the significant clusters. A procedure for merging small clusters with significant clusters is discussed by Waltman and Van Eck (2012).

Table 5A
University ranking differences according to the MNCS indicator in going from the WoS system to granularity level 8.

	First 100 universities (1)	Remaining 400 universities (2)	Total = (1) + (2)
>50 positions	2	53	55
26–50	12	101	113
16–25	13	88	101
6–15	23	97	120
≤5 positions	50	61	111
Total	100	400	500

Table 5B
University differences in MNCS values in going from the WoS system to granularity level 8.

	First 100 universities (1)	Remaining 400 universities (2)	Total = (1) + (2)
>0.20	4	3	7
>0.10 and ≤ 0.20	8	17	25
>0.05 and ≤0.10	32	97	129
≤0.05	56	283,339	
Total	100	400	500

Table 5C
Main gainers and losers in the change from the WoS system to granularity level 8 (only universities in the top 100 according to level 8 are considered) according to the MNCS indicator.

	Level 8 ranking	Re-rankings in number of positions	WoS MNCS – level 8 MNCS
Gainers	(1)	(2)	(3)
1. London School of Hygiene and Tropical Medicine	9	35	0.21
2. University of Saint Andrews	35	27	−0.09
3. University College London	39	27	−0.06
4. University of Bristol	49	26	−0.06
5. Delft University	62	36	−0.08
6. Queen Mary University London	65	62	−0.11
7. Paris Tech École Polytechnic	70	32	−0.06
8. Tech. University München	87	27	−0.04
9. University of Stuttgart	92	54	−0.08
10. Paris Diderot University	98	35	−0.06
11. McMaster University	100	28	−0.04
Losers			
1. University of Göttingen	7	6	1.78
2. Rice University	21	18	0.49
3. University Dublin Trinity College	69	46	0.21
4. University of Notre Dame	90	48	0.16
5. Lancaster University	93	36	0.11

Four comments are in order. Firstly, M_1 values along the WVE sequence are generally close to 1, which is of course the value at level 13 that consists of as many clusters as articles. In turn, M_2 values start at 1.27 at level 1, but gradually decline to 1.15 at level 12. Secondly, note that, abstracting from *ex aequo* cases, a uniform distribution would have percentages 0.50, 0.25, and 0.25 in the three categories distinguished in columns 3–5 in Table 4. However, we observe that university MNCS distributions are generally skewed, with percentages 0.56, 0.28, and 0.16 at level 1 that smoothly evolve toward 0.49, 0.31, and 0.20 at level 12. Thirdly, interestingly enough, the third effect of increasing the granularity level along the WVE sequence is a slow but continuous decrease in university MNCS inequality: as the granularity level increases, university MNCS values get closer together, as indicated by the coefficient of variation. Fourthly, mean and dispersion statistics do not clearly single out a granularity level close to the WoS system. We can perhaps point to levels 4–6 as the closest to the WoS system on this account.

Coming now to university MNCS comparisons, Tables 5A–5C in RCW presents the matrix of Pearson and Spearman correlation coefficients for all classification systems.[7] Except for level 0, Pearson correlation coefficients between university MNCS values, and Spearman correlation coefficients between ranks for any pair of classification systems, are generally high. For example, Pearson correlation coefficients between the MNCS values reached under granularity levels 4–8 and their

[7] University rankings according to all classification systems can be found in Table B in the Appendix in RCW, where universities are ordered according to the MNCS results for granularity level 8.

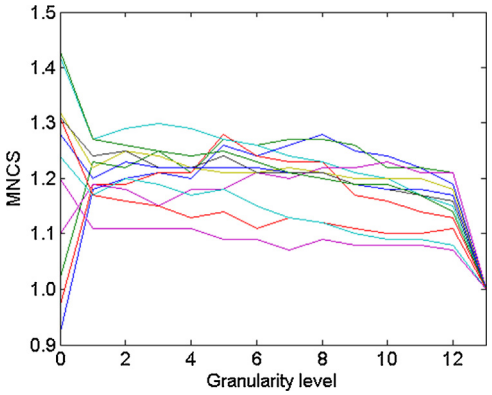

Fig. 1. Dependence of the MNCS values of 12 Dutch universities on the granularity level of the classification system.

close neighbors in the WVE sequence are equal to or greater than 0.98 in most cases. In particular, the Pearson correlation coefficient between levels 7 and 8 is 0.99. Finally, Pearson correlation coefficients between the WoS system and levels 7 and 8 are 0.96 and 0.94, respectively. Spearman correlation coefficients relating ordinal rankings are generally even higher. For example, these coefficients between the WoS system and levels 7 and 8 are both 0.97.

However, high correlations between university MNCS values and ranks do not preclude important differences for individual universities. To illustrate the sensitivity of university MNCS values to the granularity level, we use the 12 Dutch universities included in the 2013 edition of the Leiden Ranking as an example. Fig. 1 shows for each of these universities the dependence of the MNCS value on the granularity level. In addition to levels 1–12, the figure also includes level 0, at which all publications belong to the same cluster, and level 13, at which each publication belongs to its own cluster. Since at level 0 we have only one cluster, there essentially is no field normalization. Level 13 represents the other extreme. At this level, the normalization removes all differences in citation impact between publications, making citation analysis a meaningless exercise.

It is observed that moving from level 0 to level 1 brings the MNCS values of the 12 universities much closer to each other. The three universities with the lowest MNCS value at level 0 are the three technical universities in the Netherlands (i.e., Delft University of Technology, Eindhoven University of Technology, and University of Twente). These universities carry out a lot of research in engineering fields. Such fields tend to have a low citation density, which explains the low performance of the technical universities at level 0. The two universities with the highest MNCS value at level 0 are Leiden University and Erasmus University Rotterdam. The high performance of these universities at level 0 can be explained by the strong presence of the universities in biomedical fields, which tend to be fields with a high citation density. Between levels 1 and 12, it is difficult to detect general patterns. For individual universities, however, clear trends can sometimes be observed. For instance, Leiden University and Erasmus University Rotterdam both exhibit a decreasing trend. On the other hand, for two of the technical universities, a kind of inverse U-shape can be observed. Eindhoven University of Technology peaks at level 5, at which it has the highest MNCS value of all Dutch universities. Delft University of Technology has its peak at level 8, at which it also outperforms all other Dutch universities. When moving toward level 12, a weakly decreasing trend can be observed for most universities. At level 13, each university by definition has an MNCS value of exactly one.

3.2. Comparison between selected classification systems

After this illustration with Dutch universities, we turn to the comparisons between levels 7 and 8, and between level 8 and the WoS system for the entire set of 500 universities. In analyzing the consequences of going from level 7 to level 8, we must take two aspects into account. Firstly, we should analyze the re-rankings that take place in such a move. It is observed that 50% of all universities change ranks by five or fewer positions, while only 5% change ranks by more than 25 positions. Moreover, most of the large changes take place within the last 400 universities according to level 8's order. Among the first 100 universities, there is no change by more than 25 positions, and 75% of the universities experience re-rankings equal to or smaller than five positions (see Tables 6A–6C in RCW). Secondly, we should compare the differences between the university MNCS values themselves. As pointed out by Waltman, Calero-Medina, et al. (2012), since university MNCS distributions are somewhat skewed, an increase in the rank of a university by, say, 10 positions is much more significant in the top of the ranking than further down the list. Therefore, a statement such as "University X is performing 20% better than university Y according to the MNCS indicator" is more informative than a statement such as "University X is ranked 20 positions higher than university Y according to the MNCS indicator." It turns out that university differences in this respect are very small indeed: for 484 out of the 500 universities, differences in MNCS values are equal to or smaller than 0.05; in 14 other cases

Table 6A
University ranking differences according to the $PP_{top\ 10\%}$ indicator in going from the WoS system to granularity level 8.

	First 100 universities (1)	Remaining 400 universities (2)	Total = (1) + (2)
>50 positions	0	81	81
26–50	7	107	114
16–25	13	74	87
6–15	36	81	117
≤5 positions	44	57	101
Total	100	400	500

Table 6B
University differences in $PP_{top\ 10\%}$ values in going from the WoS system to granularity level 8.

	First 100 universities (1)	Remaining 400 universities (2)	WoS $PP_{top\ 10\%}$ – level 8 $PP_{top\ 10\%}$
>0.20	1	16	17
>0.10 and ≤ 0.2	12	66	78
>0.05 and ≤ 0.10	27	124	151
≤0.05	60	94	254
Total	100	400	500

Table 6C
Main gainers and losers in the change from the WoS system to granularity level 8 (only universities in the top 100 according to level 8 are considered) according to the $PP_{top\ 10\%}$ indicator.

	Level 8 ranking	Re-rankings in number of positions	WoS $PP_{top\ 10\%}$ – level 8 $PP_{top\ 10\%}$
Gainers			
1. King's College London	75	27	−0.06
2. Delft University	78	33	−0.08
3. Tech. University München	82	28	−0.07
4. University of Exeter	94	31	−0.08
5. Georgetown University	99	32	−0.07
6. University of Iowa	100	39	−0.07
Losers			
1. Rice University	23	18	0.46
2. Technical University Denmark	71	22	0.14
3. University of Notre Dame	97	26	0.11

differences are between 0.05 and 0.10; and in only two cases there is a large change going from level 7 to level 8.[8] Taking into account the small differences between levels 7 and 8, in the sequel we focus exclusively on level 8.

In the comparison between the WoS system and level 8, two aspects should be emphasized. Firstly, as observed in Table 5A, re-rankings are now more important. Only 22.2% of all universities experience small changes (at most five positions), while 33.6% change ranks by more than 25 positions. However, there are considerably fewer re-rankings among the first 100 universities (ordered according to level 8): 50 universities change ranks by five or fewer positions, while only 14 experience changes of more than 25 positions. The largest change is 62 positions. Secondly, as observed in Table 5B, differences in MNCS values are also more important than in the comparison between levels 7 and 8: 67.8% of universities experience a difference equal to or smaller than 0.05, while for 32 universities, or 6.4% of the total, the change is greater than 0.10. Interestingly enough, in 12 out of the 32 cases the latter changes take place within the first 100 universities according to level 8.

By way of example, Table 5C includes the largest gainers and losers among the first 100 universities when going from the WoS system to granularity level 8. Fourteen universities experience a re-ranking greater than 25 positions, and among the remaining 86 universities there are two that experience a change – a loss in both cases – of more than 0.20 in MNCS value. The three columns include the ranking according to level 8, the number of positions in the re-ranking, and the difference in MNCS values. Two comments are in order. Firstly, there are only three cases of universities within the first 25 in the ranking according to level 8 (London School of Hygiene and Tropical Medicine among the gainers, as well as University of Göttingen and Rice University among the losers). Together with three other gainers before the 50th rank, the remaining ten major changes when going from the WoS system to level 8 take place between position 62 and 100. Secondly, in four cases the

[8] These two cases are the University of Göttingen, a loser with a change of 1.72 – 2.29 = –0.57, and the University of Warsaw, a gainer with a change of 0.93 – 0.74 = 0.19. We note that the University of Göttingen is quite a special case. The MNCS value of the University of Göttingen is strongly determined by a single extremely highly cited publication. As a consequence, the MNCS value of this university is rather sensitive to the way in which this single publication is classified in the classification system that is used in the MNCS calculation (see Waltman et al., 2012, for more details on this case). As a further illustration of the consequences of moving from level 7 to level 8, Fig. 2 in RCW shows that Dutch universities are hardly affected by this move.

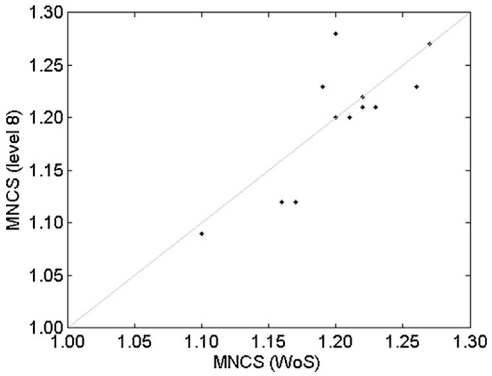

Fig. 2. Scatter plot of the MNCS values of 12 Dutch universities obtained using the WoS classification system and the classification system at granularity level 8.

differences in MNCS values are greater than 0.20 (London School of Hygiene and Tropical Medicine among the gainers, as well as University of Göttingen, Rice University, and University Dublin Trinity College among the losers).

Fig. 2 shows the differences between the WoS system and level 8 for the 12 Dutch universities. For most universities, the differences are more or less negligible. However, for some universities, more significant differences can be observed. In particular, when moving from the WoS system to level 8, Radboud University Nijmegen experiences an MNCS decrease of 0.05. On the other hand, Delft University of Technology experiences an MNCS increase of 0.08 that raises its ranking among the 12 Dutch universities from the eighth to the first position.

4. The citation impact of universities under different classification systems according to the $PP_{top\ 10\%}$ indicator

Given a classification system, percentile rank indicators directly incorporate a suitable normalization procedure for citation counts of publications from different clusters or scientific sub-fields (Bornmann & Marx, 2013). Consider, for example, the percentile rank approach in which all publications in a given scientific field are sorted out by citation numbers, and broken down into percentile ranks with values between 0 and 100. Since this procedure transforms every field citation distribution into a uniform distribution, completely eliminating the effect on citation inequality of differences in citation practices across fields, Li et al. (2013) call it a "perfect normalization" procedure that they use as a reference for the assessment of other normalization procedures. However, it is essential to understand that the "perfect normalization" offered by percentile rank indicators is conditional on the classification system that is used. There is still a need to find out which classification system is best to use.

As indicated in Section 1, in this paper we use the $PP_{top\ 10\%}$ indicator because of its prominent role in the Leiden and *SCImago* rankings. In the case of the $PP_{top\ 10\%}$ indicator, each cluster citation distribution is broken down into two sets with values 0 and 1 according to whether publications' citation counts are below or above the 90th percentile.[9] In this way, given a classification system, the $PP_{top\ 10\%}$ approach constitutes again a kind of perfect normalization procedure. The problem, of course, is that the ranking of the 500 universities still depends on which classification system we care to use.

In this scenario, this section has two aims. The first aim is to explore some general features of the variation of the citation impact of universities according to the $PP_{top\ 10\%}$ indicator under different classification systems. The second aim is to perform a direct comparison of the citation impact of universities for the WoS system and the member of the WVE sequence selected as the most convenient in the previous section, namely, level 8 (the evidence concerning the small differences between granularity levels 7 and 8 is available on request).

The information on universities' $PP_{top\ 10\%}$ values for all classification systems, as well as the country and the number of publications of each university can be found in Table C in the Appendix in RCW, where universities are ordered according to the $PP_{top\ 10\%}$ results for granularity level 8. To facilitate the comparison with the MNCS results in Table A in the Appendix in RCW, Table C reports the ratio of each university's $PP_{top\ 10\%}$ value and the world reference, namely, 10.0%. Thus, if a university has a $PP_{top\ 10\%}$ value of 11.2%, Table C in RCW reports a value of (11.2%/10.0%) = 1.12. The same way of reporting $PP_{top\ 10\%}$ values is used in the rest of this section.

With respect to the variation of the citation impact of universities under different classification systems, it should be noted that results for the $PP_{top\ 10\%}$ indicator are very similar to the results obtained for the MNCS indicator (see Table 8 in RCW for summary measures for distributions of university $PP_{top\ 10\%}$ values, and Table 9 in RCW including the matrix of Pearson and Spearman correlation coefficients for all classification systems).

[9] In fact, things are slightly more complicated, since we treat publications for which the number of citations is exactly at the top 10% threshold in a fractional way. These publications are considered to be partly in the top 10% of their field and partly in the bottom 90%. In this way, we ensure that we have exactly 10% of the publications in a field belonging to the top 10%. For more details, we refer to Waltman and Schreiber (2013).

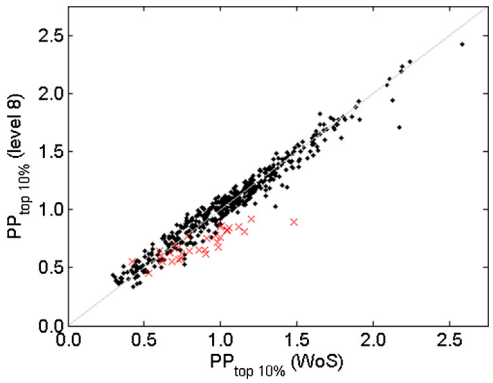

Fig. 3. Scatter plot of the $PP_{top\ 10\%}$ values of 500 universities obtained using the WoS classification system and the classification system at granularity level 8. Chinese universities (excluding Hong Kong) are indicated using a red cross. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Turning now to the comparison between selected classification systems, as in Section 3 high correlations between university $PP_{top\ 10\%}$ values and ranks do not preclude important differences for individual universities. In the key comparison between the WoS system and level 8, two aspects should be emphasized. Firstly, large re-rankings of more than 25 positions according to the $PP_{top\ 10\%}$ indicator occur somewhat more frequently than according to the MNCS indicator: 195 versus 168 universities (see Tables 5A and 6A). This is due to the situation among the last 400 universities, where these numbers are 188 and 154, respectively. Instead, among the first 100 universities according to the $PP_{top\ 10\%}$ indicator there are only seven universities with re-rankings between 26 and 50 positions and no universities with rank differences of more than 50 positions, while according to the MNCS indicator there are twelve universities with re-rankings between 26 and 50 positions, as well as two more universities with rank differences of more than 50 positions. Secondly, not surprisingly, something similar occurs when we consider differences in indicator values (see Tables 5B and 6B). Large changes greater than 0.10 in indicator values occur more or less equally frequently among the first 100 universities according to both indicators: 12 and 13 universities according to the MNCS and $PP_{top\ 10\%}$ indicators, respectively. However, these numbers become 20 and 82 among the remaining 400 universities.

By way of example, Table 6C includes the largest gainers and losers among the first 100 universities when going from the WoS system to granularity level 8. Seven universities experience a re-ranking greater than 25 positions, and among the remaining 93 universities there are two that experience a change – a loss in both cases – of more than 0.10 in $PP_{top\ 10\%}$ value. The three columns include the ranking according to level 8, the number of positions in the re-ranking, and the difference in $PP_{top\ 10\%}$ values. Note that, except for Rice University, placed in the 23rd position in level 8's ranking, all other gains and losses take place among universities placed between positions 71 and 100.

Finally, to illustrate the sensitivity of university $PP_{top\ 10\%}$ values to the choice of a classification system, we use the 32 Chinese universities (excluding Hong Kong) in the 2013 edition of the Leiden Ranking as an interesting example. Each of these universities is indicated using a red cross in Fig. 3. The main lesson is that, as we go from the WoS system to level 8, the performance of almost all Chinese universities worsens according to the $PP_{top\ 10\%}$ indicator. This deterioration is especially significant for the best performing Chinese universities.[10] An explanation of the Chinese case is beyond the scope of this paper, but it may relate to the specific characteristics of the research areas in which Chinese universities focus their activity or to the citation behavior of Chinese researchers (since researchers' citation behavior determines how publications are clustered in the WVE methodology).

5. Summary, discussion, and extensions

5.1. Summary and conclusions

In this paper, we have built a sequence of twelve classification systems by applying the WVE publication-level algorithmic methodology introduced by Waltman and Van Eck (2012) to a large WoS dataset. The dataset consists of 3.6 million publications, of the type article and review, in academic journals – excluding local journals, magazines and trade journals – published in the period 2005–2008, and with a five-year citation window. The twelve classification systems are independent, not nested or hierarchically linked, and, at each granularity level, the cluster size distribution is not restricted in any way. The reason is that we wanted to focus all attention on the consequences of adopting different granularity levels with an increasing number of clusters.

[10] The same results are obtained when we change the classification system from level 4 to level 8. For reasons of space, these results are available on request.

We have confronted two research questions. Firstly, do the characteristics of classification systems lead us to the selection of specific members in the WVE sequence to be used in the calculation of normalized citation impact indicators? Secondly, consider the possibility of evaluating the citation impact of the 500 universities in the 2013 edition of the CWTS Leiden Ranking using the MNCS and the $PP_{top\ 10\%}$ indicators. The question is: how do the results change as we change the classification system used in the evaluation exercise? In particular, how do the results change when we use the WoS classification system versus the WVE systems suggested in our answer to the first research question?

Our findings concerning the two research questions can be summarized as follows:

1. An important difference between the WoS classification system and the twelve WVE classification systems is the presence in the latter of a large number of small clusters (less than 100 publications) with a low mean citation. However, the importance of the publications included in small clusters varies dramatically across granularity levels. These publications represent less than 1% of the total for granularity levels 1–8, and more than 60% in granularity level 12.
2. As the granularity level increases, the distribution of university citation impact values according to both the MNCS and the $PP_{top\ 10\%}$ indicators gradually becomes less dispersed (according to the coefficient of variation) and less skewed (according to the CSS approach).
3. Although it is difficult to single out an optimal granularity level within the WVE sequence, we recommend the use of level 7 or 8. The percentage of articles in small clusters is still smaller than 1% of the total at these levels, and these levels clearly show a greater homogeneity than the WoS system while they capture in an acceptable way the skewness of science across clusters. Levels 7 and 8 include, respectively, 2272 and 4161 significant clusters with at least 100 publications. Hence, our analysis suggests that working with a few thousand significant clusters may be an optimal choice.
4. There is a strong correlation between the MNCS and $PP_{top\ 10\%}$ values obtained under the WoS system and most WVE granularity levels. Comparing the WoS system and granularity level 8, we obtain Pearson and Spearman correlation coefficients of 0.94 or higher. However, this does not preclude the existence of substantial differences for individual universities. For instance, when going from the WoS system to level 8, the $PP_{top\ 10\%}$ values of many Chinese universities decrease substantially.
5. In the comparison between the WoS system and level 8 using the MNCS for evaluation purposes, approximately one third of the universities change ranks by more than 25 positions. Also, almost one third of the universities experience a difference in MNCS values greater than 0.05. Of these universities, there are seven for which the difference is even above 0.20.
6. Differences are somewhat more important when using the $PP_{top\ 10\%}$ indicator: 39% of all universities change ranks by more than 25 positions, while almost half of the universities experience a difference in $PP_{top\ 10\%}$ values greater than 0.05. There are 17 universities for which this difference is above 0.20. However, among the last 400 universities relatively large differences are more frequent than among the first 100. As a matter of fact, large differences between the WoS system and level 8 among the first 100 universities are more prevalent when using the MNCS indicator than when using the $PP_{top\ 10\%}$ indicator.

5.2. Discussion

Performing an accurate correction for field-specific factors is far from trivial. In general, field normalization requires specifying an adequate classification system. This is a problem for which there is no perfect solution. In practice, fields do not have clear-cut boundaries. Fields tend to overlap, and their boundaries tend to be fuzzy. Moreover, fields can be defined at many different levels of aggregation, and it is unclear which level is most appropriate for the purpose of normalizing citation impact indicators. Given these difficulties, Kostoff and Martinez (2005) even conclude that a "*. . .meaningful* '*discipline*' *citation average may not exist, and the mainstream large-scale mass production semi-automated citation analysis comparisons may provide questionable results*." (op. cit, p. 61).

Consequently, it must be recognized that any field normalization of citation impact indicators involves a certain degree of arbitrariness caused by the methodology used to define fields. In this scenario, we have developed a proposal for a normalization approach that is likely to be more accurate than the approach based on the well-known WoS classification system. In so doing, we have also provided some insight into the sensitivity of citation impact indicators to the choice of a classification system.

Our findings lead to the following two remarks. Firstly, for the purpose of field normalization, we believe that our algorithmically constructed classification systems offer an attractive alternative to the WoS classification system. Unlike the WoS system, our algorithmically constructed systems are defined at the level of individual publications rather than at the level of entire journals. Our systems are therefore better able to handle publications in multidisciplinary journals and in other journals with a broad scope. Furthermore, our algorithmically constructed systems can be expected to offer an up-to-date representation of the structure of scientific fields. This may not always be the case for the WoS system. Based on the criteria we have developed, having between 2000 and 4000 significant clusters with more than 100 publications in an algorithmically constructed classification system seems to be a good choice. However, it should be recognized that working with algorithmically constructed classification systems poses a troublesome labeling problem (Waltman & Van Eck, 2012) that, in certain contexts, may limit its applicability. In addition, one should be aware that at a high aggregation level Waltman and

Van Eck (2012) report only a partial correspondence between research areas in algorithmically constructed classification systems and traditional disciplines such as chemistry, computer science, engineering, mathematics, etc.

Secondly, consider the application of the MNCS and $PP_{top\,10\%}$ indicators at the level of universities. For certain general analytical purposes, some readers may conclude that the consequences of choosing between the WoS system and level 8 for normalization purposes (see Tables 5A–5C and 6A–6C) turn out to be relatively small for most universities, above all for the first 100 universities. However, it should be recognized that differences of 0.05 or more affect from one third to one half of all universities. The problem is that in practice there is often a tendency to pay serious attention even to rather small differences in the values of a citation impact indicator. Our results show that this introduces a significant risk of over-interpretation. For instance, in the case of both indicators applied at the university level, differences of 0.05 may well relate to the choice of a certain classification system and may therefore have little meaning in terms of actual differences in the citation impact of the universities' publications.

5.3. Extensions

As suggestions for future research, we would like to mention five possible extensions of our work.

1. An interesting possibility would be to compare classification systems not only at the level of science as a whole but also at lower levels, for instance at the level of a number of broad disciplines. For this purpose, one could use the broad disciplines obtained from an algorithmically constructed classification system at a low granularity level. At the disciplinary level, differences between classification systems can be expected to have a more significant effect than at the level of science as a whole.
2. It would be important to investigate whether there are any systematic factors that help explain the apparition of gainers and losers as the granularity level increases. For instance, the remarkable results for Chinese universities deserve further investigation.
3. Given an algorithmically constructed classification system, such as level 8 in this paper, one could confront it with other interesting available alternatives that have been used to challenge the WoS system. For example, one could classify articles into the 80 sections distinguished in Chemical Abstracts for Chemistry and related fields (Neuhaus & Daniel, 2009), or into the 19 specialties distinguished in EconLit for the field of Economics (Van Leeuwen & Calero-Medina, 2012), and study the clusters where they are classified in level 8. Among other things, this exercise may help in the labeling problem of algorithmically constructed classification systems.
4. Another possible extension would be to compare our approach with two recently proposed approaches. Firstly, users of WoS databases typically use the WoS journal subject categories as building blocks for field normalization. As we know, this requires the prior solution of the assignment problem of publications belonging to two or more subject categories, which in this paper has been solved following a fractional approach. Alternatively, Rons (2012) develops a so-called partition-based field normalization that uses as building blocks the smaller cells of the partition created by the WoS subject categories and their intersections. In this way, as in the WVE methodology, every publication is assigned to a single building block. Secondly, Colliander (2014) suggests the idea of identifying for each publication a set of related publications. The citation impact of a publication can then be determined by comparing the number of citations received by the publication with the number of times related publications have been cited.
5. Finally, our research could be extended by considering the use of algorithmically constructed classification systems in which publications are allowed to belong to multiple clusters. This probably offers an improved way of dealing with publications that are of an interdisciplinary nature. Techniques that could potentially be used to construct classification systems with overlapping clusters have been proposed in the recent network science literature (Ball, Karrer, & Newman, 2011; Gopalan & Blei, 2013).

Acknowledgements

This paper was conceived while Ruiz-Castillo enjoyed the hospitality of the Centre for Science and Technology Studies, The Netherlands, during the 2013 spring term. Ruiz-Castillo also acknowledges financial help from the Spanish MEC through grant ECO2011-29762. Suggestions by two referees have helped us to improve the final version of the paper.

References

Adams, J., Gurney, K., & Jackson, L. (2008). Calibrating the zoom – A test of Zitt's hypothesis. *Scientometrics*, *75*, 81–95.

Albarrán, P., & Ruiz-Castillo, J. (2011). References made and citations received by scientific articles. *Journal of the American Society for Information Science and Technology*, *62*, 40–49.

Albarrán, P., Crespo, J., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, *88*, 385–397.

Ball, B., Karrer, B., & Newman, M. E. J. (2011). Efficient and principled method for detecting communities in networks. *Physical Review E*, *84*, 036103.

Bornmann, L., & Marx, W. (2013). How good is research really? *EMBO Reports*, *14*, 226–230.

Colliander, C. (2014). A novel approach to citation normalization: A similarity-based method for creating reference sets. *Journal of the Association for Information Science and Technology*, http://dx.doi.org/10.1002/asi.23193 (in press)

Colliander, C., & Ahlgren, P. (2011). The effects and their stability of field normalization baseline on relative performance with respect to citation impact: A case study of 20 natural science departments. *Journal of Informetrics*, *5*, 101–113.

Crespo, J. A., Li, Y., & Ruiz-Castillo, J. (2013). The measurement of the effect on citation inequality of differences in citation practices across scientific fields. *PLoS ONE*, *8*, e58727.
Crespo, J. A., Herranz, N., Li, Y., & Ruiz-Castillo, J. (2014). The effect on citation inequality of differences in citation practices at the Web of Science subject category level. *Journal of the American Society for Information Science and Technology*, *65*, 1244–1256.
Glänzel, W. (2007). Characteristic scores and scales: A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics*, *1*, 92–102.
Glänzel, W. (2011). The application of characteristic scores and scales to the evaluation and ranking of scientific journals. *Journal of Information Science*, *37*, 40–48.
Glänzel, W., Thijs, B., Schubert, A., & Debackere, K. (2009). Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics*, *78*, 165–188.
Gopalan, P. K., & Blei, D. M. (2013). Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, *110*, 14534–14539.
Kostoff, R. N., & Martinez, W. L. (2005). Is citation normalization realistic? *Journal of Information Science*, *31*, 57–61.
Leydesdorff, L., & Bornmann, L. (2014). The operationalization of fields as WoS subject categories (WCs) in evaluative bibliometrics: The cases of Library and Information Science and Science & Technology Studies. *Journal of the Association for Information Science and Technology*,
Li, Y., Castellano, C., Radicchi, F., & Ruiz-Castillo, J. (2013). Quantitative evaluation of alternative field normalization procedures. *Journal of Informetrics*, *7*, 746–755.
Neuhaus, C., & Daniel, H.-D. (2009). A new reference standard for citation analysis in chemistry and related fields based on the sections of chemical abstract. *Scientometrics*, *78*, 219–229.
de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, *149*, 510–515.
Radicchi, F., & Castellano, C. (2012). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS ONE*, *7*, e33833.
Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences United States of America*, *105*, 17268–17272.
Rons, N. (2012). Partition-based field normalization: An approach to highly specialized publication records. *Journal of Informetrics*, *6*, 1–10.
Ruiz-Castillo, J. (2014). The comparison of classification-system-based normalization procedures with source normalization alternatives in Waltman and Van Eck. *Journal of Informetrics*, *8*, 25–28.
Ruiz-Castillo, J., & Costas, R. (2014). The skewness of scientific productivity. *Journal of Informetrics*, *8*, 917–934.
Ruiz-Castillo, J., & Waltman, L. (2014). *Field-normalized citation impact indicators using algorithmically constructed classification systems of science. Working paper 14-03*. Universidad Carlos III. http://hdl.handle.net/10016/18385
Schubert, A., Glänzel, W., & Braun, T. (1987). Subject field characteristic citation scores and scales for assessing research performance. *Scientometrics*, *12*, 267–292.
Seglen, P. (1992). The skewness of science. *Journal of the American Society for Information Science*, *43*, 628–638.
Van Eck, N. J., Waltman, L., Van Raan, A. F. J., Klautz, R. J. M., & Peul, W. C. (2013). Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PLoS ONE*, *8*, e62395.
Van Leeuwen, T. N., & Calero-Medina, C. (2012). Redefining the field of economics: Improving field normalization for the application of bibliometric techniques in the field of economics. *Research Evaluation*, *21*, 61–70.
Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology*, *64*, 372–379.
Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*, 2378–2392.
Waltman, L., & Van Eck, N. J. (2013a). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison. *Scientometrics*, *96*, 699–716.
Waltman, L., & Van Eck, N. J. (2013b). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, *7*, 833–849.
Waltman, L., & Van Eck, N. J. (2013c). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B*, *86*, 471.
Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., Visser, M. S., & Van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*, 37–47.
Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., Van Eck, N. J., et al. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*, 2419–2432.
Waltman, L., Van Eck, N. J., & Van Raan, A. F. J. (2012). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, *63*, 72–77.
Zitt, M., Ramana-Rahari, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalization. *Scientometrics*, *63*, 373–401.