Perianes-Rodriguez, A.; Ruiz-Castillo, J. (2017). A comparison of the Web of Science and publication-level classification systems of science. Journal of Infonometrics, v. 11, n. 1, pp. 32-45. Available in https://doi.org/10.1016/j.joi.2016.10.007.

A comparison of the Web of Science and publication-level classification systems of science

Antonio Perianes-Rodriguez [a], Javier Ruiz-Castillo [b,*]

[a] *Departamento de Biblioteconomía y Documentación, Universidad Carlos III, SCImago Research Group, Spain*

[b] *Departamento de Economía, Universidad Carlos III, Spain*

Keywords:
Classification systems of science
Journal-level *versus* publication-level systems
Field-normalization

ABSTRACT

In this paper, we propose a new criterion for choosing between a pair of classification systems of science that assign publications (or journals) to a set of clusters. Consider the standard target (cited-side) normalization procedure in which cluster mean citations are used as normalization factors. We recommend system A over system B whenever the standard normalization procedure based on system A performs better than the standard normalization procedure based on system B. Performance is assessed in terms of two double tests – one graphical, and one numerical – that use both classification systems for evaluation purposes. In addition, a pair of classification systems is compared using a third, independent classification system for evaluation purposes. We illustrate this strategy by comparing a Web of Science journal-level classification system, consisting of 236 journal subject categories, with two publication-level algorithmically constructed classification systems consisting of 1363 and 5119 clusters. There are two main findings. Firstly, the second publication-level system is found to dominate the first. Secondly, the publication-level system at the highest granularity level and the Web of Science journal-level system are found to be non-comparable. Nevertheless, we find reasons to recommend the publication-level option.

1. Introduction

For many theoretical and practical purposes in the evaluation of research activities in current society, we need a *classification system* of science, that is, an assignment of individual publications (or journals) to a set of clusters or sub-fields. As is well known, the choice of a classification system remains an open question in Scientometrics (see *inter alia* Boyack, Klavans, Börner, 2005 ; Leydesdorff, 2004, 2006; Small, 1999; Leydersdorff & Rafols, 2009, as well as the references they contain). Together with the classification systems included in Thomson Reuters' Web of Science (WoS) and Elsevier's Scopus databases, there are a number of interesting proposals suggested by individual researchers (see *inter alia* Börner et al. (2012), as well as the references in Waltman & Van Eck, 2012).[1]

In this paper, we contribute to the search for an appropriate classification system begun in Ruiz-Castillo and Waltman (2015). The main idea is the following. Given a classification system, it is well known that differences in production and

* Corresponding author.
E-mail address: jrc@eco.uc3m.es (J. Ruiz-Castillo).

[1] The historical background section of Klavans and Boyack (2015) contains an illuminating guide to the literature on the construction of "research fronts" and publication-level or journal-level "taxonomies" (or classification systems).

citation practices preclude the direct comparison of the raw citations received by any pair of publications belonging to different clusters. In this situation, one way to evaluate the performance of research units working in different clusters begins with the normalization of the original citation counts. Consider the standard target (or cited-side) normalization procedure in which normalized citation scores in every cluster are equal to the original raw citations divided by the cluster mean citation. If one could establish that the standard normalization procedure based on system A performs better than the standard normalization procedure based on system B, then we would recommend the use of system A over system B. In this paper, we use the graphical and numerical methods introduced in Li and Ruiz-Castillo (2013) for that purpose. Following up Sirtes (2012) and Waltman & Van Eck (2013a), these methods include the possibility of using a third, independent classification system C for the evaluation of any pair of systems A and B.

We illustrate this strategy by comparing a Web of Science (WoS) journal-level classification system, consisting of 236 journal subject categories (or simply categories hereafter), with two alternatives arising from the publication-level algorithmic methodology introduced in Waltman & Van Eck (2012) that classifies individual publications into clusters solely based on direct citations between them.

In practice, the choice of the WoS classification system is often made because, together with the Scopus system, it is readily available. However, a number of studies question the appropriateness of this system for normalization purposes.[2] Among the publication-level alternatives, Klavans and Boyack (2015) conclude that classification systems based on direct citation using the Waltman & Van Eck (2012) methodology are more accurate than classification systems based on bibliographic coupling or co-citation. Ruiz-Castillo and Waltman (2015) apply the publication-level algorithmic methodology introduced by Waltman & Van Eck (2012) to a WoS dataset consisting of 9.4 million publications from the 2003–2012 period. They construct a sequence of twelve independent classification systems, in each of which the same set of publications is assigned to an increasing number of clusters. In this paper, we use the versions obtained at granularity levels 6 and 8 (the G6 and G8 classification systems hereafter) consisting of 1363 and 5119 clusters, respectively. Therefore, we have three standard normalization procedures based on three classification systems, and two interesting comparisons to make: the G6 versus the G8 system, and the winner in this contest versus the WoS system.

We focus on the 3.6 million articles published in the 2005–2008 period, and the citations they receive during a five-year citation window for each year in that period. However, two complications should be noted. Firstly, approximately 45% of the articles in the WoS system are assigned to two or more categories up to a maximum of six. To deal with this problem, we adopt a multiplicative strategy in which articles classified into several categories are wholly counted in all of them. In this way, the space of articles is expanded as much as necessary beyond the initial size. Secondly, since the methods for the evaluation of normalization procedures in Li and Ruiz-Castillo (2013) require the partition of cluster (and category) citation distributions into, say, 100 quantiles, we eliminate clusters (and categories) with less than 250 articles.

The remainder of this paper consists of four Sections. Section II presents the data. Section III serves two purposes: the description of the graphical and numerical methods for the comparison of the performance of two normalization procedures based on two different classification systems, and the application of these methods to the comparison between the G6- and G8-normalization procedures. Since the G8 system performs better than the G6 system, Section IV compares the performance of the WoS- and the G8-normalization procedures. Finally, Section V discusses the results and offers some concluding comments. The Supplementary material (SM hereafter) includes some descriptive statistics, a numerical example illustrating the various citation distributions used in the paper, and a method to evaluate the differences between a pair of classification systems in different circumstances.

2. Data

Our dataset results from the application of a publication-level algorithmic methodology to 9,446,622 distinct articles published in 2003–2012. Publications in local journals, as well as popular magazines and trade journals, have been excluded (for details, see Ruiz-Castillo and Waltman, 2015). 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.

In this paper, we focus on the set of 3,614,447 distinct articles published in the period 2005–2008, and the 31,290,249 citations received by these articles during a five-year citation window for each year in that period. The percentage of distinct articles assigned to two or more categories is very high: 45.2% of the total (for details, see Section A in the SM).[3] To deal with the problem of multiple assignment of articles to WoS categories, we adopt a multiplicative strategy in which articles classified into several sub-fields are wholly counted in all of them. In this way, the space of articles is expanded as much as necessary beyond the initial size. As a matter of fact, the total number of articles in what we call the *extended count* for the 236 WoS categories is 5,944,533, or 64.5% larger than the original dataset. The number of citations in the extended count is 50,307,834, or 60.8% more than in the original dataset.

[2] See *inter alia* Neuhaus and Daniel (2009) for Chemistry and related fields, Van Leeuwen and Calero-Medina (2012) for Economics & Business, Van Eck, Waltman, Van Raan, Klautz, and Peul (2013) for Clinical and Basic Medical Research, and Leydesdorff and Bornmann (2015) and Wang and Waltman, 2016 for Library and Information Science, and Science & Technology Studies.

[3] This amount is of the same order as that found in other comparable datasets. For example, this percentage is 42% in the WoS dataset of 3.7 million articles published in the 1998–2002 period that was used in Albarrán, Crespo, Ortuño, and Ruiz-Castillo, 2011a.

The G6 system and, above all, the G8 system, are plagued with small clusters with typically low mean citation rates. In the G8 system, after the elimination of 34.9% of clusters with less than 250 articles, we are left with 95.2% of all articles – classified into 3332 clusters –, and 97.6% of all citations in the original dataset. Similarly, in the G6 system, after the elimination of 34.0% small clusters, we are left with 9.7% of the total number of distinct articles – classified into 900 clusters –, and 99.9% of the total number of citations.

For comparison purposes, we also eliminate the five WoS categories with less than 250 articles. In this case we are left with more than 99.9% of all articles and citations relative to the initial extended count.

3. The comparison between the G6 and G8 classification systems

3.1. Preliminaries

Although it is difficult to single out an optimal granularity level within the sequence of twelve classification systems studied in Ruiz-Castillo and Waltman (2015), these authors make the following considerations.

(i) An important difference between the WoS system and the twelve publication-level systems is the presence in the latter of a large number of small clusters with less than 100 articles in a four-year publication period from 2005 to 2008. The percentage of articles in small clusters varies dramatically across granularity levels. Since in levels 9–12 this percentage increases from 3.2% to 61.3% of the total, a granularity level below level 9 is recommended.

(ii) When we restrict the analysis to significant clusters with at least 100 publications, the high skewness and the similarity across cluster citation distributions usually found in the literature are well captured at all granularity levels.

(iii) If we choose a granularity level dominated by a relatively small number of broad fields, the danger is that some of them are too heterogeneous, in which case comparisons between publications within the same field may be biased. Under the assumption that as the granularity level and the number of clusters increase the degree of within-cluster homogeneity also increases, Ruiz-Castillo and Waltman (2015) use an additively decomposable citation inequality index to approximate the degree of homogeneity at every granularity level. To achieve at least a comparable degree of homogeneity as the WoS system itself, one should use a granularity level equal to or greater than level 6.

These arguments led Ruiz-Castillo and Waltman (2015) to recommend the use of classification systems with a few thousand significant clusters with at least 100 publications, such as granularity levels G7 and G8 with 2272 and 4161 significant clusters. However, in order to magnify the differences between acceptable granularity levels, in this paper we choose to analyze the G8 and G6 classification systems, where the latter has 952 significant clusters.

Point (iii) above warrants the following comment. The greater the within-cluster homogeneity is, the greater the heterogeneity across clusters will be, that is, the between-cluster differences in production and citation practices. This opens the way to the possibility of choosing between granularity levels using a new criterion. Given a granularity level, consider an adequate target (cited-side) normalization procedure for facilitating the comparability between citation counts in clusters characterized by different production and citation practices. The new criterion consists of choosing between granularity levels by comparing the performance of their target normalization procedures.

As far as target normalization procedures are concerned, Li, Castellano, Radicchi, and Ruiz-Castillo (2013) establish that the best alternative is the two-parameter system developed in Radicci and Castellano (2012). However, different results indicate that the standard, one-parameter normalization procedure, in which normalized citation scores in every field are equal to the original raw citations divided by the field mean citation, exhibits a good performance (Crespo, Li, & Ruiz-Castillo, 2013, Crespo, Herranz, Li, & Ruiz-Castillo, 2014; Li et al., 2013; Perianes-Rodriguez & Ruiz-Castillo, 2015; Radicchi, Fortunato, & Castellano, 2008; Ruiz-Castillo, 2014). Given its simplicity and good performance, in this paper we adopt this procedure for all classification systems.

On the other hand, the comparison of normalization procedures based on a single classification system in Li et al. (2013) and Perianes-Rodriguez and Ruiz-Castillo (2015) uses the measuring framework introduced in Crespo et al. (2013). In turn, the comparison of standard normalization procedures based on different classification systems needed in this paper will be made using the extension of this framework developed in Li and Ruiz-Castillo (2013).

3.2. Notation

Let $\boldsymbol{c_j}$ be the citation distribution of cluster j in system G8, ordered according to the relationship $\leq$, where $j = 1,\ldots, 3332$. The union $\boldsymbol{C} = \cup_j \{\boldsymbol{c_j}\}$ is the overall citation distribution for the set of distinct articles in the G8 system. Similarly, let $\boldsymbol{d_g}$ be the ordered citation distribution of cluster g in system G6, where $g = 1,\ldots, 900$, and let $\boldsymbol{D} = \cup_g \{\boldsymbol{d_g}\}$ be the overall citation distribution in this case. Finally, let $\boldsymbol{e_k}$ be the ordered extended citation distribution of category k, where $k = 1,\ldots, 231$. The union $\boldsymbol{E} = \cup_k \{\boldsymbol{e_k}\}$ is the overall citation distribution for the extended set of articles in the WoS system. As already indicated, $|\boldsymbol{C}| = 3.4$, $|\boldsymbol{D}| = 3.6$, and $|\boldsymbol{E}| = 5.9$ million articles.

Let $\boldsymbol{c^*_j}$ be the normalized citation distribution of cluster j in system G8, where the raw number of citations received by each article is divided by the mean citation in distribution $\boldsymbol{c_j}$, say μ_j. The union $\boldsymbol{C^*} = \cup_j \{\boldsymbol{c^*_j}\}$ is the overall G8-normalized citation distribution. In the G8 system, $|\boldsymbol{C}| = |\boldsymbol{C^*}| = 3.4$ million articles. Similarly, let $\boldsymbol{d^*_g}$ be the normalized citation distribution

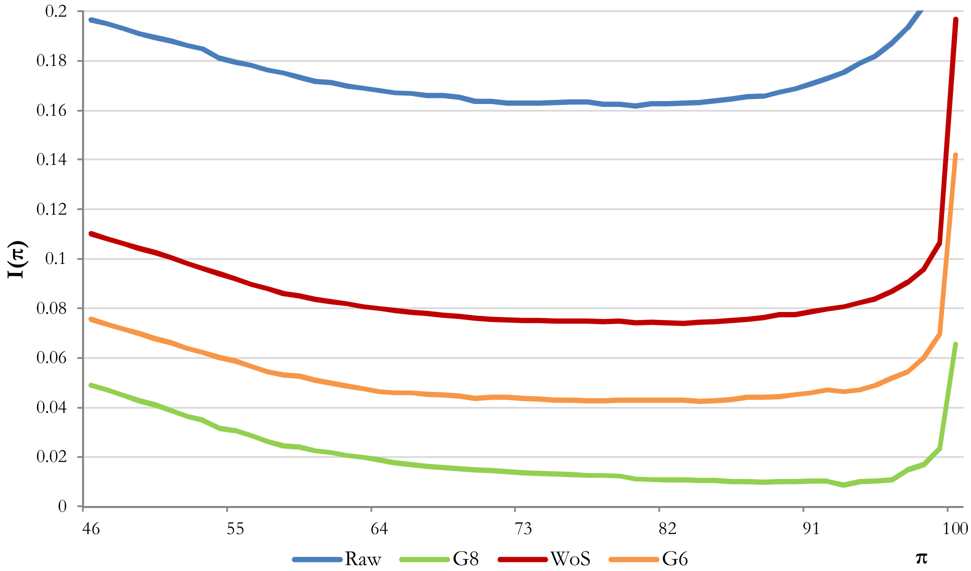

Fig. 1. $I(\pi)$ terms for percentiles in the interval [46,100] before and after normalization using the G8 classification system for evaluation purposes. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

of cluster g in system G6, where the raw number of citations received by each article is divided by the mean citation in distribution $\boldsymbol{d_g}$, say M_g. Finally, let the union $\boldsymbol{D^*} = \cup_g \{\boldsymbol{d^*_g}\}$ be the overall G6-normalized citation distribution. In the G6 system, $|\boldsymbol{D}| = |\boldsymbol{D^*}| = 3.6$ million articles.

3.3. The graphical approach to the comparison of two classification systems

In this sub-section, we evaluate the performance of a pair of normalization procedures using a method that applies the additive decomposability property of a certain member of the Generalized Entropy family of citation inequality indices – denoted by I – to the double partition of the data into clusters and quantiles (see Crespo et al., 2013; for details). Consider, for example, the case of the G8 system. Partition each citation distribution c_j into Π quantiles of equal size, $c^{\pi}{}_j$, indexed by $\pi = 1,\ldots, \Pi$. In practice, we take the percentiles, so that $\Pi = 100$. Assume for a moment that, in any cluster j we disregard the citation inequality within every percentile π by assigning to every article in that percentile the mean citation of the percentile itself, $\mu_j{}^{\pi}$.

For any π, all quantities $\mu_j{}^{\boldsymbol{\pi}}, j = 1,\ldots J$, are comparable because they represent the mean citation of publications belonging top the same percentile π in the corresponding citation distribution c_j. Thus, the interpretation of the fact that, for example, $\mu_j{}^{\pi} = 2\,\mu_l{}^{\pi}$ is that, on average, the citation impact of cluster j is twice as large as the citation impact of cluster l in spite of the fact that both quantities represent a common underlying phenomenon, namely, the same *degree* π *of citation impact or* (*citation*) *excellence* in both clusters. In other words, for any π, the difference between $\mu_j{}^{\pi}$ and $\mu_l{}^{\pi}$ is entirely attributable to the differences in the production and citation practices that prevail in the two clusters for publications having the same degree of excellence. Thus, the citation inequality between clusters at each percentile, $I(\pi) = I(\mu_1{}^{\pi},\ldots, \mu_j{}^{\pi},\ldots, \mu_J{}^{\pi})$, is entirely attributable to the differences in citation practices between the 3332 clusters, holding constant the degree of excellence in all clusters at percentile π.

Consequently, to assess the impact of the G8-normalization procedure using the G8 system for evaluation purposes, we simply observe how expressions $I(\pi)$ vary when we compute them for the normalized citation distributions $\boldsymbol{c^*_j}$, $j = 1,\ldots$, 3332. The two alternatives, before and after normalization, correspond to the blue and the green lines in Fig. 1 (Since the terms $I(\pi)$ are very high for percentiles in the lower tail of citation distributions, for clarity Fig. 1 only includes percentiles π in the interval [46,100]). Note that the impact of the G8-normalization procedure is very important: the green line is considerably below the blue line at all percentiles.

For the comparison of the G8- and G6-normalization procedures, we extend the methods introduced in Li and Ruiz-Castillo (2013) to take into account that the citation distributions $\boldsymbol{C}$ and $\boldsymbol{D}$ have a different number of articles. We begin by assessing the performance of the G6-normalization procedure when the G8 system is used for evaluation purposes. To do this, consider the articles belonging to both systems, that is, consider the set $\boldsymbol{C} \cap \boldsymbol{D}$. For every distinct article i in this set receiving c_{ji} citations in the cluster citation distribution $\boldsymbol{c_j}$ in system G8, there must be one article u in some cluster g in the G6 system with the same number of citations, i. e. $c_{ji} = d_{gu}$. The normalized score of this article in the G8 system becomes $c^{**}{}_{ji} = c_{ji}/M_g$, where M_g is the mean citation in cluster g. In this way, we construct an overall G6-normalized citation distribution under the G8 system $\boldsymbol{C^{**}} = \cup_j \{\boldsymbol{c^{**}_j}\}$. Since $c_{ji}/M_g = d_{gu}/M_g = d^*{}_{gu}$, so that $c^{**}{}_{ji} = d^*{}_{gu}$, every article in $\boldsymbol{C^{**}}$ is contained in the citation distribution $\boldsymbol{D^*}$. However, since there are some articles in the G6 system that belong to a small cluster with

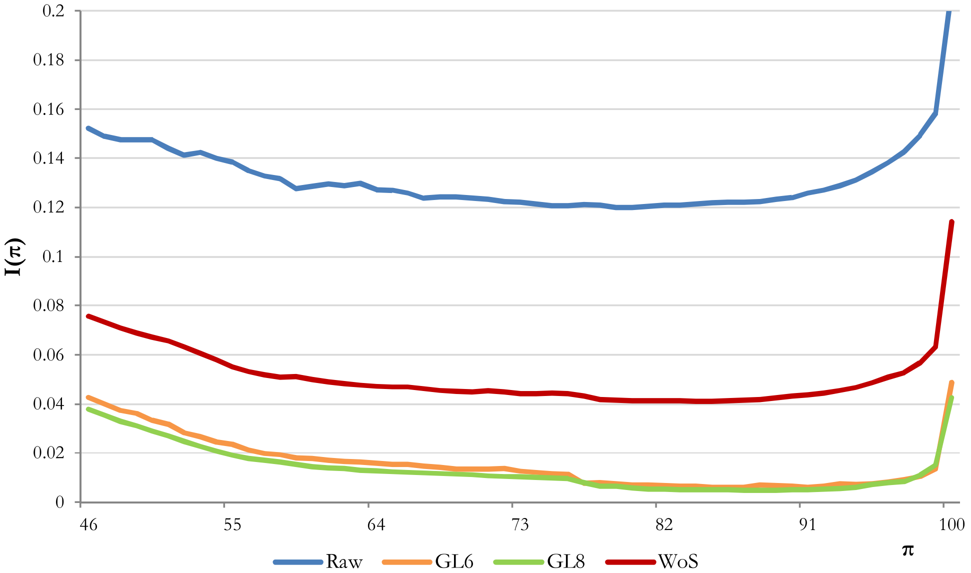

Fig. 2. $I(\pi)$ terms for percentiles in the interval [46,100] before and after normalization using the G6 classification system for evaluation purposes. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

less than 250 articles in the original G8 system, we have that $\boldsymbol{C}^{**} \subset \boldsymbol{D}^{*}$, so that $|\boldsymbol{C}^{**}| < |\boldsymbol{D}^{*}| = 3.6$ million articles. It is useful to construct a numerical example to illustrate this procedure. However, to facilitate the reading of the text the example is included in the SM. Sections B.1 and B.2 in the SM describe simplified versions of the G8 and the G6 classification systems, while Section B.3 illustrates the construction of the $\boldsymbol{C}^{**}$ citation distribution.

The orange line in Fig. 1 represents the expressions $I(\pi)$ for the G6-normalization procedure when the G8 system is used for evaluation purposes, that is, for the organization of the data in terms of $\boldsymbol{C}^{**}$. The fact that the orange line is above the green line indicates that the effect of the G6-normalization procedure in reducing $I(\pi)$ at every π is not as strong as the effect of the G8-based normalization procedure. In this situation, we say that the latter *uniformly dominates the former in the graphical approach using the G8 classification system for evaluation purposes*.

It has been argued that the assessment of two normalization procedures would be generally biased in favor of the normalization procedure based on the classification system used for evaluation purposes (Sirtes, 2012; Waltman & Van Eck, 2013a,b). According to Li and Ruiz-Castillo (2013), a solution consists of adding a second test to the above procedure where the G6 system is now used for evaluation purposes. The partition of each citation distribution $\boldsymbol{d_g}$ into percentiles $\boldsymbol{d}\pi_{\boldsymbol{g}}$ for all π and g, gives rise to a set of $I(\pi)$ expressions before normalization. To assess the impact of the G6-based normalization procedure we simply observe how expressions $I(\pi)$ vary when we compute them for the normalized citation distributions $\boldsymbol{d}^{*}_{\boldsymbol{g}}$, $g = 1, \ldots, 900$. The two alternatives, before and after normalization, correspond to the blue and the orange lines in Fig. 2 (as before, for clarity Fig. 2 only includes percentiles π in the interval [46,100]). Again, the impact of the G6-normalization procedure is very important: the orange line is always clearly below the blue line for all π.

Denote by $\boldsymbol{D}^{**} = \cup_g \{\boldsymbol{d}^{**}_{\boldsymbol{g}}\}$ the overall G8-normalized citation distribution under the G6 system, whose construction is explained in Section B.4 in the SM. The green line in Fig. 2 represents the expressions $I(\pi)$ for the G8-based normalization procedure when the G6 system is used for evaluation purposes, that is, for the organization of the data in terms of $\boldsymbol{D}^{**}$. Since the green and the orange line intersect at some percentiles, we cannot say that one system uniformly dominates the other in the graphical approach. Thus, in this case we say that the G8- and the G6-normalization procedures are *non-comparable in the graphical approach when using the G6 system for evaluation purposes*. However, since the former uniformly dominates the latter using G8 as the evaluation classification system, using the terminology of Li and Ruiz-Castillo (2013) we conclude that the G8-normalization procedure *weakly dominates* the G6-normalization procedure *according to the double test in the graphical approach*.

3.4. The numerical approach to the comparison of two classification systems

The comparison of two classification systems is done in terms of the performance of the standard field-normalization procedures based on them. As we have seen, the performance of the two field-normalization procedures in the graphical approach is assessed in terms of the lines representing expressions $I(\pi)$ for all π after field-normalization in each case. This has the advantage that no value judgment is used to weight differences between two expressions $I(\pi)$ and $I(\pi')$ at different percentiles, say π and π'. However, the comparison of entire lines is somewhat cumbersome. More importantly, as long as the two lines intersect, the graphical approach leads to the non-comparability of the procedures in question. As we will presently see, the numerical approach comes to the rescue –albeit at the cost of introducing a new value judgment.

Table 1
The impact of the G8-, G6-, and WoS-normalization procedures using the G8 classification system for evaluation purposes.

Overall citation inequality due to differences in citation practices, raw data	Overall citation inequality due to differences in citation practices after normalization		
	According to G8:	According to G6:	According to WoS:
$IDCP/I(C)$ 22.7% Expressions: (2)	$IDCP^*/I(C^*)$ 4.6% (3)	$IDCP(G6)^*/I(C^{**})$ 9.3% (4)	$IDCP(WoS)^*/I(C^{***})$ 13.6% (17)
Impact of the G8-normalization procedure $[IDCP - IDCP^*]/IDCP$ 83.4% Expressions (5)	Impact of the G6-normalization procedure $[IDCP - IDCP(\mathrm{G6})^*]/IDCP$ 64.3% (6)		Impact of the WoS-normalization procedure $[IDCP - IDCP(\mathrm{WoS})^*]/IDCP$ 46.6% (18)

In the context of the G8 system, Crespo et al. (2013) propose a numerical estimate of the effect on overall citation inequality, $I(\boldsymbol{C})$, which can be attributed to differences in production and citation practices between the 3332 clusters through a term denoted *IDCP* (*I*nequality due to *D*ifferences in *C*itation *P*ractices). It can be shown that $I(\boldsymbol{C})$ can be expressed as the sum of three terms, one of which is the *IDCP* term defined as follows:

$$IDCP = \Sigma_{\pi} \mathrm{v}^{\pi} I(\pi) \tag{1}$$

where, for each π, $\boldsymbol{v}^{\boldsymbol{\pi}}$ is the share of total citations received by articles in quantiles $\boldsymbol{c}^{\boldsymbol{\pi}}{}_j$ for all j, so that Σ_{π} $\boldsymbol{v}^{\boldsymbol{\pi}} = 1$. Therefore, the term *IDCP* is a weighted average of the quantities $I(\pi)$, with weights $\boldsymbol{v}^{\boldsymbol{\pi}}$ that add up to one. It should be noted that, due to the skewness of science, in practical applications the weights $\boldsymbol{v}^{\boldsymbol{\pi}}$ tend to increase dramatically with π. For assessing the relative effect on the overall citation inequality $I(\boldsymbol{C})$ attributed to the differences in production and citation practices between the 3332 clusters in system G8, we use the ratio

$$IDCP/I(\mathrm{C}) \tag{2}$$

Therefore, in the numerical approach we use a single number, $IDCP/I(\boldsymbol{C})$, to summarize such a relative effect at every level of excellence, i. e. at all $\pi = 1,\ldots, 100$. The cost of this simplification is that we have used a very specific weighting scheme, $\boldsymbol{v}^{\boldsymbol{\pi}}$ for each π, for representing the entire set of expressions $I(\pi)$, $\pi = 1,\ldots, 100$, into a single number $IDCP = \Sigma_{\pi}$ $\boldsymbol{v}^{\boldsymbol{\pi}}$ $I(\pi)$.

On the other hand, let $IDCP^*$ and $IDCP(\mathrm{G6})^*$ be the *IDCP* terms after applying the G8- and G6-normalization procedures using the G8 system for evaluation purposes. Since $I(\boldsymbol{C^*})$ and $I(\boldsymbol{C^{**}})$ are the overall citation inequality after applying the two normalization procedures, their impact can be assessed by the ratios

$$IDCP^*/I(C^*), \tag{3}$$

and

$$IDCP(\mathrm{G6})^*/I(C^{**}). \tag{4}$$

In order to compare the performance of the two procedures it is useful to measure the relative change in the *IDCP* term in both cases, namely, the ratios

$$[IDCP - IDCP^*]/IDCP, \tag{5}$$

and

$$[IDCP - IDCP(\mathrm{G6})^*]/IDCP. \tag{6}$$

The results for expressions 2 to 6 are presented in Table 1.

Differences in production and citation practices between the 3332 clusters in the G8 system are responsible for 22.7% (expression 2 in Table 1) of overall citation inequality $I(\boldsymbol{C})$. In agreement with the graphical approach (green *versus* blue line in Fig. 1), the G8-normalization procedure considerably reduces this percentage to 4.6% (expression 3 in Table 1). Instead, the G6-normalization procedure using the G8 system for evaluation purposes reduces this percentage only to 9.3% (orange *versus* blue line in Fig. 1, and expression 4 in Table 1). For comparison purposes, we observe that the G8- and G6-normalization procedures reduce the *IDCP* term by 83.4% and 64.3%, respectively (expressions 5 and 6 in Table 1), i.e. the former clearly *dominates* the latter *in the numerical approach when the G8 system is used for evaluation purposes*.

We now turn to the analysis of the effect on overall citation inequality that can be attributed to differences in production and citation practices between the 900 clusters in the G6 system, measured by a term denoted *IDCP*'. If we let $IDCP'^*$ and $IDCP(\mathrm{G8})'^*$ measure the effect on overall citation inequality attributed to the differences in cluster citation distributions after applying the G6- and the G8-normalization procedures using the G6 system for evaluation purposes, then expressions (2) to (4) become:

$$IDCP'/I(\mathrm{D}), \tag{7}$$

Table 2
The impact of the G8-, G6-, and WoS-normalization procedures using the G6 classification system for evaluation purposes.

Overall citation inequality due to differences in citation practices, raw data	Overall citation inequality due to differences in citation practices after normalization		
	According to G6:	According to G8:	According to WoS:
IDCP'/*I*(*D*) 17.0% Expressions: (7)	*IDCP* */*I*(*D**) 3.49% (8)	*IDCP*(G8)'*/*I*(*D***) 3.45% (9)	*IDCP(WoS) */I(D***)* 8.5% (21)
Impact of the G8-normalization procedure [*IDCP*' – *IDCP*(G8) *]/*IDCP*' 83.6% Expressions (10)	Impact of the G6-normalization procedure [*IDCP*' – *IDCP*'*]/*IDCP*' 82.1% (11)		Impact of the WoS-normalization procedure [*IDCP*' – *IDCP*(*WoS*)'*]/*IDCP*' 55.5% (22)

$$IDCP'^{*}/I(\mathrm{D}^{*}), \tag{8}$$

$$\mathrm{and}IDCP(\mathrm{G8})'^{*}/I(\mathrm{D}^{**}), \tag{9}$$

where $I(\boldsymbol{D})$, $I(\boldsymbol{D^*})$, and $I(\boldsymbol{D^{**}})$ are the overall citation inequality before and after applying the two normalization procedures. In turn, the relative change in the *IDCP*' term in both cases is given by the ratios

$$[IDCP' - IDCP'^{*}]/IDCP', \tag{10}$$

$$[IDCP' - IDCP(G8)'^{*}]/IDCP'. \tag{11}$$

The results for expressions 7 to 11 are presented in Table 2.

Differences in production and citation practices between the 900 clusters in the G6 system are now responsible for 17.0% (expression 7 in Table 2) of overall citation inequality $I(\boldsymbol{D})$ – an amount smaller than the corresponding figure in the case of 3332 clusters. Under the G6 system, the G6-normalization procedure considerably reduces this percentage to 3.49% (orange *versus* blue line in Fig. 2, and expression 8 in Table 2), whereas the G8-normalization procedure reduces this percentage down to 3.45% (green *versus* blue line in Fig. 2, and expression 9 in Table 2). Finally, we observe that the G6- and G8-normalization procedures lower the *IDCP* term by 82.1% and 83.6% (expressions 10 and 11 in Table 2), i.e. the latter barely *dominates* the former *in the numerical approach when the G6 system is used for evaluation purposes*. This illustrates how the comparison between the G6 and the G8 systems using G6 for evaluation purposes, that was impossible under the graphical approach, now can be resolved in the numerical approach.

We conclude that the G8-normalization procedure *dominates* the G6-normalization procedure *according to the double test in the numerical approach*.

3.5. *Robustness analysis*

Ideally, for comparing two normalization procedures based on two different classification systems we should use a third, independent system, for evaluation purposes (Sirtes, 2012; Waltman & Van Eck, 2013a). As originally pointed out by Zitt, Ramana-Rahari, and Bassecoulard (2005) in the context of classification systems at different aggregation levels, an outstanding article in a certain sub-field may get only a modest score in a larger field if the rest of the articles in the latter have more generous referencing practices. Following Zitt et al. (2005), we consider the possibility of computing the set of the top *X*% most cited publications in every cluster in a pair of classification systems A and B. An article that belongs to the top *X*% in cluster *j* in system A may not belong to the top *X*% in cluster *l* in system B. The more often this is the case, the more different the two systems will be according to the *X*% criterion. To compare the WoS system with systems G8 and G6 we must take into account that the comparison can be made in terms of distinct or extended articles. The extension of the original method can be found in Section C in SM. The results for the comparison between the WoS and the G8 systems and the WoS and the G6 systems for values of *X* equal to 50%, 10%, and 1% are presented in Table 3a and b, respectively.

Since the extended count is greater, differences in terms of extended articles are always larger than differences in terms of distinct publications. Two points should be noted. Firstly, differences between the G8 and the WoS systems are somewhat larger for any *X* than differences between the G6 and the WoS systems. For example, the difference between the two first classification systems in the top 1% of most cited articles is, approximately, 50% (last column in row V in Table 3a). Secondly, at least in the upper tails of clusters' and categories' citation distributions, the systems G6 and G8 are quite different from the WoS system. Therefore, we suggest comparing the G6- and G8-normalization procedures using the WoS classification system for evaluation purposes.

In Section III.2 we introduced the overall citation distribution for the extended set of articles in the WoS system, $\boldsymbol{E} = \cup_k \{\boldsymbol{e_k}\}$. Let $\boldsymbol{e^*_k}$ be the normalized extended citation distribution of category k, where the raw number of citations received by each article is divided by the mean citation in distribution $\boldsymbol{e_k}$, say m_k. The union $\boldsymbol{E^*} = \cup_k \{\boldsymbol{e^*_k}\}$ is the overall WoS-normalized citation distribution. Section B.5 in the SM describes simplified versions of distributions $\boldsymbol{E}$ and $\boldsymbol{E^*}$ in the example. In order to assess the performance of the G8- and G6-normalization procedure when the WoS system is used for evaluation purposes,

Table 3
(a) Differences in% between the top most cited articles in the G8 and WoS classification systems. (b) Differences in% between the top most cited articles in the G6 and WoS classification systems. (c) Differences in% between the top most cited articles in the G6 and G8 classification systems.

	Most cited articles		
	Top 50%	Top 10%	Top 1%
V. Difference in distinct articles	9.2%	25.2%	47.3%
VI. Difference in extended articles	15.6%	33.6%	53.2%

	Most cited articles		
	Top 50%	Top 10%	Top 1%
III. Difference in distinct articles	8.5%	21.7%	40.1%
IV. Difference in extended articles	11.0%	27.6%	47.8%

	Most cited articles		
	Top 50%	Top 10%	Top 1%
I. Difference in terms of the G6 system	12.3%	22.2%	33.2%
II. Difference in terms of the G8 system	8.3%	18.9%	32.4%

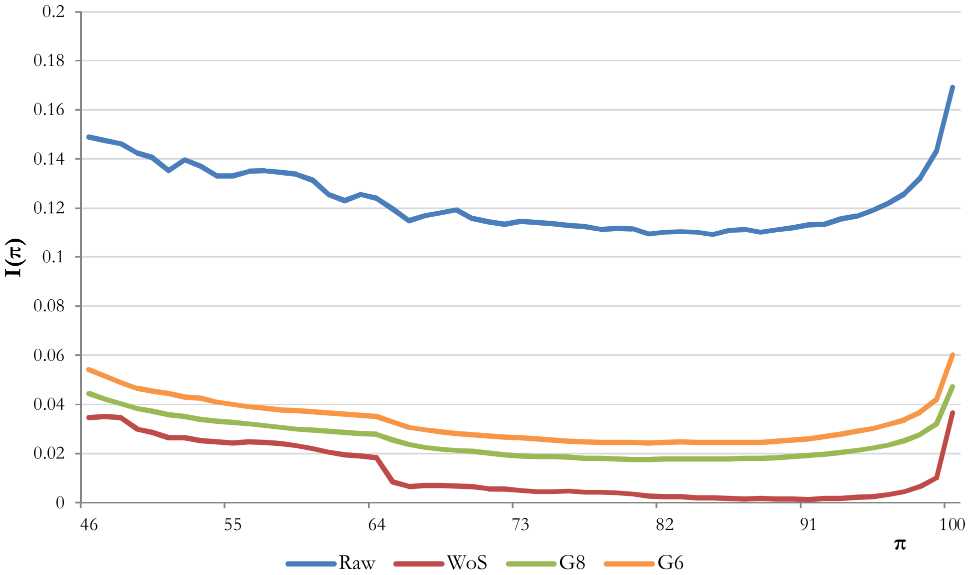

Fig. 3. $I(\pi)$ terms for percentiles in the interval [46,100] before and after normalization using the WoS classification system for evaluation purposes. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

let $\boldsymbol{E}^{**} = \cup_k \{\boldsymbol{e}^{**}_{\boldsymbol{k}}\}$ and $\boldsymbol{E}^{***} = \cup_k \{\boldsymbol{e}^{***}_{\boldsymbol{k}}\}$ be the G8- and G6-normalized overall citation distributions under the WoS system. Sections B.6 and B.7 in SM illustrate the construction of the $\boldsymbol{E}^{**}$ and $\boldsymbol{E}^{***}$ citation distributions.

The results of the comparison between the G6- and G8-normalization procedures under the graphical approach using the WoS system for evaluation purposes are in Fig. 3 (as before, Fig. 3 only includes percentiles π in the interval [46,100]). The expressions $I(\pi)$ corresponding to the organization of the raw data according to $\boldsymbol{E} = \cup_k \{\boldsymbol{e_k}\}$ are represented by the blue line in Fig. 3. On the other hand, the orange and the green lines represent the effect of the G6- and the G8-normalization procedures when the data is organized according to $\boldsymbol{E}^{***}$ and $\boldsymbol{E}^{**}$, respectively. Clearly, we conclude that the G8 system *strictly dominates* the G6 system *in the graphical approach using the WoS classification system for evaluation purposes*.

We now turn to the robustness analysis under the numerical approach. The effect on overall citation inequality that can be attributed to differences in production and citation practices between the 231 categories in the WoS system is measured by a term denoted *IDCP*". We use the ratio

$$IDCP''/I(\mathrm{E}) \tag{12}$$

for assessing the relative effect of *IDCP*" on the overall citation inequality $I(\boldsymbol{E})$. If we let *IDCP*(G8)"* and *IDCP*(G6)"* measure the effect on overall citation inequality attributed to the differences in categories citation distributions after applying the G8-

Table 4
The impact of the G8-, G6-, and WoS-normalization procedures using the WoS classification system for evaluation purposes.

Overall citation inequality due to differences in citation practices, raw data	Overall citation inequality due to differences in citation practices after normalization		
	According to G8:	According to G6:	According to WoS:
$IDCP''/I(E)$ 15.4% Expressions: (12)	$IDCP(\mathrm{G8})''^{*}/I(E^{**})$ 5.2% (13)	$IDCP(\mathrm{G6})''^{*}/I(E^{***})$ 5.81% (14)	$IDCP''^{*}/I(E^{*})$ 2.8% (19)
Impact of the G8-normalization procedure $[IDCP'' - IDCP(\mathrm{G8})''^{*}]/IDCP''$ 73.0% Expressions: (15)	Impact of the G6-normalization procedure $[IDCP'' - IDCP(\mathrm{G6})''^{*}]/IDCP''$ 67.2% (16)		Impact of the WoS-normalization procedure $[IDCP'' - IDCP''^{*}]/IDCP''$ 84.0% (20)

and the G6-normalization procedures using the WoS system for evaluation purposes, then the impact of the two procedures is given by the expressions

$$IDCP(\mathrm{G8})''^{*}/I(\mathrm{E}^{**}), \qquad (13)$$

and

$$IDCP(\mathrm{G6})''^{*}/I(\mathrm{E}^{***}), \qquad (14)$$

where $I(\boldsymbol{E}^{**})$, and $I(\boldsymbol{E}^{***})$ are the overall citation inequality after applying the two normalization procedures. In turn, the relative change in the *IDCP*” term in both cases is given by the ratios

$$[IDCP'' - IDCP(\mathrm{G8})''^{*}]/IDCP'' \qquad (15)$$

and

$$[IDCP'' - IDCP(\mathrm{G6})''^{*}]/IDCP''. \qquad (16)$$

The results for expressions 12 to 16 are presented in Table 4.

Differences in production and citation practices between the 231 categories in the WoS classification system are now responsible for 15.4% (expression 12 in Table 4) of overall citation inequality $I(\boldsymbol{E})$. In agreement with the graphical approach, the G6-normalization procedure considerably reduces this percentage to 5.8% (orange *versus* blue line in Fig. 3, and expression 14 in Table 4), whereas the G8-normalization procedure reduces this percentage down to 5.2% (green *versus* blue line in Fig. 3, and expression 13 in Table 4). Finally, we observe that the G6- and G8-normalization procedures lower the *IDCP* term by 67.2% and 73.0% (expressions 16 and 15 in Table 4), i.e. the latter slightly dominates the former under the numerical approach when the WoS system is used for evaluation purposes.

4. The comparison between the G8 and WoS classification systems

4.1. The graphical approach

In order to assess the performance of the WoS-normalization procedure, we begin by constructing the WoS-normalized overall citation distribution under the G8 system, $\boldsymbol{C}^{***} = \cup_j \{\boldsymbol{c}^{***}_{\boldsymbol{j}}\}$ (see Section B.8 in SM). The red line in Fig. 1 represents the expressions $I(\pi)$ for the WoS-based normalization procedure when the G8 system is used for evaluation purposes, that is, for the organization of the data in terms of $\boldsymbol{C}^{***}$. The fact that the red line is above the green line indicates that the effect of the WoS-based normalization procedure in reducing $I(\pi)$ at every π is not as strong as the effect of the G8-based normalization procedure. We say that the latter *uniformly dominates* the former *in the graphical approach using G8 as the evaluation classification system*.

Next, we must assess the performance of the G8- and the WoS-normalization procedures using the WoS system for evaluation purposes. To assess the impact of the WoS-based normalization procedure we simply observe how expressions $I(\pi)$ vary when we compute them for the normalized citation distributions $\boldsymbol{e}^{*}_{\boldsymbol{k}}$, $k = 1, \ldots, 231$. The two alternatives, before and after normalization, correspond to the blue and the red lines in Fig. 3. It is observed that the impact of this normalization is very important: the red line is always clearly below the blue line for all π. In turn, as we have seen in Section III.2, the green line in Fig. 3 represents the expressions $I(\pi)$ for the G8-normalization procedure when the WoS system is used for evaluation purposes, that is, for the organization of the data in terms of $\boldsymbol{E}^{**}$. Since the red line is below the green line for all π, the effect of the WoS-normalization procedure in reducing $I(\pi)$ at every π is stronger than the effect of the G8-normalization procedure. We say that the former *uniformly dominates* the latter *in the graphical approach using the WoS classification system for evaluation purposes*.

We conclude that the two normalization procedures are *non-comparable in terms of the double test in the graphical approach*. Nevertheless, insofar as in Fig. 3 the distance between the green and the red lines is smaller than in Fig. 1, we may say that the

G8-normalization procedure performs better using the WoS system for evaluation purposes than the WoS-normalization procedure using the G8 system for evaluation purposes.

4.2. The numerical approach

Let $IDCP(\mathrm{WoS})^*$ be the $IDCP$ term after applying the WoS-normalization procedure using the G8 system for evaluation purposes. The impact of this procedure can be assessed by the ratio

$$IDCP(\mathrm{WoS})^*/I(\mathrm{C}^{***}). \tag{17}$$

For comparative purposes, we assess the performance of this procedure in terms of the relative change in the $IDCP$ term, namely, the ratio

$$[IDCP - IDCP(\mathrm{WoS})^*]/IDCP. \tag{18}$$

The results for expressions 17 and 18 are presented in Table 1.

As we have seen in Section III.4, differences in production and citation practices between the 3332 clusters in system G8 are responsible for 22.7% of overall citation inequality $I(\boldsymbol{C})$, while the G8-normalization procedure considerably reduces this percentage to 4.6%. Instead, the WoS-normalization procedure using the G8 system for evaluation purposes lowers this percentage only to 13.6% (red *versus* blue line in Fig. 1, and expression 17 in Table 1). For comparison purposes, we observe that the G8- and WoS-normalization procedures reduce the $IDCP$ term by 83.4% and 46.6% (expressions 5 and 18 in Table 1), i.e. the former clearly *dominates* the latter *in the numerical approach when we use the G8 system for evaluation purposes*.

We now turn to the analysis of the effect on overall citation inequality that can be attributed to differences in production and citation practices between the 231 categories in the WoS system, measured by the term $IDCP$". If $IDCP"^*$ measures the effect on overall citation inequality attributed to the differences in category citation distributions after applying the WoS-normalization procedure, then the impact of this procedure can be measured by the expression

$$IDCP'' * /I(\mathrm{E}^*), \tag{19}$$

Whereas for comparative purposes the performance of this procedure can be assessed in terms of the relative change in the $IDCP$" term:

$$[IDCP'' - IDCP''^*]/IDCP''. \tag{20}$$

The results for expressions 19 and 20 are presented in Table 4.

According to expression 12 in Table 4, differences in production and citation practices between the 231 categories in the WoS system are now responsible for 15.4% of overall citation inequality $I(\boldsymbol{E})$. The role of the G8- and WoS-normalization procedures is reversed: the WoS-normalization procedure considerably lowers this percentage to 2.8% (red *versus* blue line in Fig. 3), whereas the G8-normalization procedure reduces this percentage only to 5.2% (green *versus* blue line in Fig. 3). On the other hand, the WoS- and G8-normalization procedures reduce the $IDCP$ term by 84.0% and 73.0%, i.e. the former *dominates* the latter *in the numerical approach when the WoS system is used for evaluation purposes*.

As in the graphical approach, we confirm that the superiority of the G8-normalization procedure over the WoS-normalization procedure using the G8 system for evaluation purposes under the numerical approach (expressions 5 and 18 in Table 3) is greater than the superiority of the WoS-normalization procedure over the G8-normalization procedure using the WoS system for evaluation purposes (expressions 20 and 15 in Table 4).[4]

4.3. Robustness analysis

As we know, for comparing two normalization procedures based on two different classification systems ideally we should use a third, independent system, for evaluation purposes. In our case, the G6 system cannot be considered fully independent from the G8 system because both have been constructed with the same algorithmic methodology. However, we should recall that the classification systems in Ruiz-Castillo and Waltman (2015) are not hierarchically linked: by fixing the resolution parameter at twelve different values, a sequence of independent classification systems is built, in each of which the same set of publications is assigned to an increasing number of clusters. In any case, although differences between the WoS and the G6 systems (Table 3b) are greater than differences between the G8 and the G6 systems (Table 3c), the latter are by no means negligible.[5] Consequently, we believe that it is useful to compare the G8- and WoS-normalization procedures using the G6 classification system for evaluation purposes.

[4] Although the order of magnitude is smaller, the same result is obtained when we compare the G6 and the WoS systems: the superiority of the G6-normalization procedure over the WoS-normalization procedure using the G6 system for evaluation purposes under the numerical approach (expression 11 *versus* expression 22 in Table 2) is greater than the superiority of the WoS-normalization procedure over the G6-normalization procedure using the WoS system for evaluation purposes (expression 20 *versus* expression 16 in Table 4).

[5] This is the same result obtained in Zitt et al. (2005) for a WoS dataset with five different aggregation levels.

Denote by $\boldsymbol{D}^{***} = \cup_g \{\boldsymbol{d}^{***}{}_{\boldsymbol{g}}\}$ the WoS-normalized overall citation distribution under the G6 system, whose construction is described in Section B.9 in SM. As we saw in Section III.2, the green line in Fig. 2 represents the effect of the G8-normalization procedure using the G6 system for evaluation purposes. In turn, the red line in Fig. 2 represents the effect of the WoS-normalization procedure using the G6 system for evaluation purposes, that is, when the data is organized according to $\boldsymbol{D}^{***}$. The fact that the red line is always above the green line in Fig. 2 indicates that the G8-normalization procedure *strongly dominates* the WoS-normalization procedure *in the graphical approach when using G6 as the evaluation classification system*.

For the numerical analysis, the impact of the G8-normalization procedure is measured in expressions 9 and 10 in Table 2. As far the impact of the WoS-normalization procedure, consider the expressions

$$IDCP(\mathrm{WoS})^{*}/I(\mathrm{D}^{***}). \tag{21}$$

and

$$[IDCP - IDCP(\mathrm{WoS})^{*}]/IDCP, \tag{22}$$

included in Table 2.

As we saw in expression 7 in Table 2, differences in production and citation practices between the 900 clusters in the G6 system are responsible for 17.0% of overall citation inequality $I(\boldsymbol{D})$. The G8-normalization procedure reduces this percentage to 3.45% (expression 9 in Table 2). In contrast, the WoS-normalization procedure lowers this percentage only to 8.5% (expression 21 in Table 2). For comparison purposes, we observe that the G8- and WoS-normalization procedures reduce the *IDCP* term by 83.6% and 55.5% (expressions 10 and 22 in Table 2), i.e. under the numerical approach, the G8-normalization procedure *dominates* the WoS-normalization procedure *in the numerical approach when the G6 system is used for evaluation purposes*.[6]

5. Discussion and conclusions

5.1. Summary of the approach

Ideally, one would like to use classification systems in which the citation impact of articles in the same cluster is directly comparable. In other words, one would like to work with classification systems with a high degree of within-cluster homogeneity. However, as pointed out in Section III.1, the greater the within-cluster homogeneity is, the greater the heterogeneity across clusters will be, that is, the between-cluster differences in production and selection practices. Consequently, in this paper we have used a new criterion for choosing between a pair of classification systems: we should use system A over system B whenever the standard normalization procedure based on system A performs better according to the graphical and numerical approaches than the standard normalization procedure based on system B. The two approaches are based on a double test that uses both classification systems for evaluation purposes. In addition, a pair of classification systems can be compared using a third, independent classification system for evaluation purposes.

These ideas have been applied in two cases: the choice between publication-level algorithmically constructed classification systems G6 and G8 with 900 and 3332 clusters with at least 250 articles in a four-year publication period, and the choice between the winner in this contest, namely, system G8, and the WoS system consisting of 231 categories with at least 250 articles in the same period.

5.2. Main results

As expected, the greater the number of clusters/categories, the greater the effect on overall citation inequality attributable to differences in production and citation practices between clusters/categories. However, when the evaluation is done in terms of their own classification system, the standard normalization procedures based on the G6, G8, and WoS systems perform similarly well in the three cases (after normalization the corresponding *IDCP* terms are typically reduced by, approximately, 83%). When the comparison between normalization procedures recognizes that they are based on different systems, the main results are the following two.

1. The possibility that using a classification system for evaluation purposes biases the analysis in favor of the normalization procedure based on this system, makes it very difficult to conclude that one system-based normalization procedure surpasses another according to the double tests in the graphical and the numerical approaches. This is why the following finding is remarkable: system G8 dominates system G6 in the weak and the strong sense in the graphical and the numerical approach, respectively.[7] In addition, when the WoS system is used for evaluation purposes, the G8 system graphically and numerically dominates system G6.

[6] Although the order of magnitude of the differences is smaller than before, it is worth pointing out that the G6 system dominates the WoS system using the G8 system for evaluation purposes, both in the graphical approach (orange and red lines in Fig. 1), and in the numerical approach (expressions 6 and 18 in Table 1).

[7] This is exactly the same finding obtained in Li and Ruiz-Castillo (2013) when they compared two hierarchically nested WoS classification systems consisting of 219 subject categories and 19 broad fields.

These results have important practical consequences. Firstly, when we have a choice between two classification systems at different granularity levels, we should use the system at the higher level because it typically exhibits a better standard normalization performance when cluster mean citations are used as normalization factors. Secondly, the G6-normalization procedure has been found to perform well not only under the G6 system itself, but also when its performance is assessed using the G8 or the WoS systems for evaluation purposes (in which case it reduces the corresponding IDCP terms by 64.3% and 67.2%, respectively). Therefore, if there is only available a single classification system at an appropriately high granularity level, we should use it in the knowledge that the reduction of the effect on overall citation inequality attributable to differences in production and citation practices – even at higher granularity levels – is non-negligible.

2. As Ruiz-Castillo and Waltman (2015) recognize, the choice of an adequate classification system constitutes a quandary for which there is no perfect solution: all options involve a certain degree of arbitrariness in the way clusters are selected. Nevertheless, using a set of new gold standards – consisting of articles with at least 100 references –, Klavans and Boyack (2015) compare publication-level algorithmically constructed classification systems based on direct citations, à la Waltman & Van Eck (2012), with six journal-level systems that do not include the WoS. They conclude that the former are more accurate than the latter in the sense that they are better at concentrating references. Furthermore, it can be argued that publication-level systems are better able to handle publications in multidisciplinary journals and in other journals with a broad scope, and can be expected to offer an up-to-date representation of the structure of scientific fields (Ruiz-Castillo and Waltman, 2015). On the other hand, it should be recognized that algorithmically constructed classification systems at sufficiently high granularity levels pose a troublesome labeling problem that, in certain contexts, may limit their applicability.

In this context, this paper has compared the G8- and WoS-based standard normalization procedures. The main result is that, according to the double tests in the graphical and the numerical approaches, the two procedures are non-comparable. Nevertheless, according to both approaches, the G8-normalization procedure performs better using the WoS system for evaluation purposes than the WoS-normalization procedure using the G8 system for evaluation purposes. Furthermore, when we use the G6 system for evaluation purposes, the G8-normalization procedure performs better than the WoS-normalization procedure in the graphical and numerical sense.[8]

We conclude that these options constitute a credible alternative to the WoS system and, by extension, to other journal-based classification systems. Consequently, we celebrate the decision by the Centre for Science and Technology Studies of adopting an algorithmically constructed classification system of this type as of the 2015 edition of the Leiden Ranking.

5.3. Limitations and further research

Before we finish, we must discuss the following three questions relating to the limitations of the approach followed in this paper.

Firstly, we have focused on the possibility of choosing between classification systems depending on the performance of the standard normalization procedures based on each of the two contending systems. However, there are other non-standard, target (cited-side) normalization procedures whose performance could be equally tested with the same purpose (consider, for example, the procedures studied in Li et al., 2013).

Secondly, there are other approaches to the normalization problem for which the results of this paper are quite irrelevant. This is the case of source (or citing-side) normalization procedures in which citation weights are functions of the citing papers independently of any classification system (see *inter alia* Waltman & Van Eck, 2013b; and the references cited there).

Thirdly, percentile rank indicators directly incorporate a suitable normalization procedure for citation counts of publications from different clusters or scientific sub-fields (see *inter alia* Bornmann and Marx, 2013; and the references cited there). The percentile rank approach transforms cluster citation distributions in a way that completely eliminates the effect on citation inequality of differences in production and citation practices between clusters.[9] Similarly, consider the evaluation of research units using a high-impact citation indicator defined on the set of publications with citations above a certain high-impact threshold. Assuming that the indicator is scale- and size-independent (where the size of a citation distribution is measured by its number of publications), a research unit's performance in the all-sciences case can be evaluated using an appropriate weighted average of the unit's citation impact in each cluster (Perianes-Rodríguez and Ruiz-Castillo, 2015). However, it should be noted that both the percentile rank approach and the use of scale- and size-independent high-impact indicators are still conditional on the classification system used. Consider, for example, the *Top X%* citation impact indicator of scientific excellence defined as the percentage of an institution's scientific output included in the set formed by the *X%* of the most cited papers in every scientific cluster. Obviously, to compute this indicator we need a prior assignment of publications to clusters, i.e. a classification system. Ruiz-Castillo and Waltman (2015) studied the consequences of using the WoS and the G8 classification systems for the ranking of the 500 universities in the 2013 edition of the Leiden Ranking according to the

[8] As indicated in footnotes 4 and 6, although the order of magnitude of the differences is smaller, the same conclusions are obtained in the comparison between the G6 and the WoS systems.

[9] Consider, for example, the possibility 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.

Top 10% indicator. In this paper, we have seen that differences between these two classification systems in the top 10% most cited articles range from 25.2% to 33.6%. However, these differences range from 47.3% and 53.2% in the top 1% most cited articles (Table 3a). Therefore, we expect that the choice between classification systems could have dramatic consequences for the ranking of research units when using high-impact indicators defined over the very upper tail of citation distributions. We leave this conjecture for further research.

Author contributions

Antonio Perianes-Rodriguez: Collected the data, performed the analysis.
Javier Ruiz-Castillo: Conceived and designed the analysis, wrote the paper.

Acknowledgements

This research project builds on earlier work started by Antonio Perianes-Rodriguez during a research visit to the Centre for Science and Technology Studies (CWTS) of Leiden University as awardee of José Castillejo grant, CAS15/00178, funded by the Spanish MEC. Ruiz-Castillo is a visiting researcher at CWTS and gratefully acknowledges CWTS for the use of its data. Ruiz-Castillo acknowledges financial support from the Spanish MEC through grant ECO2014-55953-P, as well as grant MDM 2014-0431 to his Departamento de Economía. Comments from two anonymous referees have helped to improve the paper. Any remaining shortcomings are the authors' sole responsibility.

References

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.

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., et al. (2012). Design and update of a classification system: The UCSD map of science. *Public Library of Science*, *7*(7), e39464.

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

Boyack, K., Klavans, R., & Börner, K. (2005). Mapping the backbone of Science. *Scientometrics*, *64*, 351–374.

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. *Public Library of Science*, *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.

Klavans, R., & Boyack, K. W. (2015). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? arXiv:1511. 05078v1 [cs. DL]. *Journal of the Association for Information Science and Technology*,

Leydesdorff, L., & Bornmann, L. (2015). 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*, http://dx.doi.org/10.1002/asi.23408

Leydesdorff, L. (2004). Top-down decomposition of the journal citation report of the social science citation index: Graph- and factor analytical approaches. *Scientometrics*, *60*, 159–180.

Leydesdorff, L. (2006). Can scientific journals Be classified in terms of aggregated journal-Journal citation relations using the journal citation reports? *Journal of the American Society for Information Science and Technology*, *57*, 601–613.

Li, Y., & Ruiz-Castillo, J. (2013). The comparison of normalization procedures based on different classification systems. *Journal of Informetrics*, *7*, 945–958.

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. *Scientometrics*, *78*, 219–229.

Perianes-Rodríguez, A., & Ruiz-Castillo, J. (2015). A comparison of two ways of evaluating research units working in different scientific fields. *Scientometrics*, *106*, 539–561.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *PNAS*, *105*, 17268–17272.

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, *9*, 102–117.

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.

Sirtes, D. (2012). Finding the easter eggs hidden by oneself: Why Radicchi and Castellano's (2012) fairness test for citation indicators is not fair. *Journal of Informetrics*, *6*, 448–450.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, *50*, 799–813.

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. *Public Library of Science*, *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., & 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). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, *7*, 833–849.

Waltman, L., & Van Eck, N. J. (2013b). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison. *Scientometrics*, *96*, 699–716.

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, http://dx.doi.org/10.1016/j.joi.2016.02.003

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.