An Alternative to Field-normalization in the Aggregation of Heterogeneous Scientific Fields

Antonio Perianes-Rodriguez[1] and Javier Ruiz-Castillo[2]

[1] *antonio.perianes@uc3m.es*
Universidad Carlos III, Department of Library and Information Science, SCImago Research Group, C/ Madrid, 128, 28903 Getafe, Madrid (Spain)

[2] *jrc@eco.uc3m.es*
Universidad Carlos III, Departamento de Economía, C/ Madrid, 126, 28903 Getafe, Madrid (Spain)

Abstract
A possible solution to the problem of aggregating heterogeneous fields in the all-sciences case relies on the normalization of the raw citations received by all publications. In this paper, we study an alternative solution that does not require any citation normalization. Provided one uses size- and scale-independent indicators, the citation impact of any research unit can be calculated as the average (weighted by the publication output) of the citation impact that the unit achieves in all fields. The two alternatives are confronted when the research output of the 500 universities in the 2013 edition of the CWTS Leiden Ranking is evaluated using two citation impact indicators with very different properties. We use a large Web of Science dataset consisting of 3.6 million articles published in the 2005-2008 period, and a classification system distinguishing between 5,119 clusters. The main two findings are as follows. Firstly, differences in production and citation practices between the 3,332 clusters with more than 250 publications account for 22.5% of the overall citation inequality. After the standard field-normalization procedure where cluster mean citations are used as normalization factors, this figure is reduced to 4.3%. Secondly, the differences between the university rankings according to the two solutions for the all-sciences aggregation problem are of a small order of magnitude for both citation impact indicators.

Conference Topic
Indicators; Citation and co-citation analysis

Introduction
As is well known, the comparison of the citation impact of research units is plagued with obstacles of all sorts. For our purposes in this paper, it is useful to distinguish between the following three basic difficulties. (i) How can we compare the citation distributions of research units of different sizes even if they work in the same homogeneous scientific field? For example, how can we compare the output of the large Economics department at Harvard University with the output of the relatively small Economics department at Johns Hopkins? The next two difficulties have to do with the heterogeneity of scientific fields: the well-known differences in production and citation practices makes it impossible to directly compare the raw citations received by articles belonging to different fields. Given a classification system, that is, a rule for assigning any set of articles to a number of scientific fields, field heterogeneity presents the following classic hindrances in the evaluation of research units' performance. (ii) How can we compare the citation impact of two research units working in different fields? For example, how can we compare the citation impact of MIT in Organic Chemistry with the citation impact of Oxford University in Statistics and Probability? Finally, (iii) how can we compare the citation impact of two research units taking into account their

output in all fields? For example, how can we compare the citation impact of MIT and Oxford University in what we call the *all-sciences* case?

As is well known, the solution to the first two problems requires size- and scale-independent citation impact indicators. We will refer to indicators with these two properties as *admissible* indicators. Given an admissible indicator, in this paper we are concerned with the two types of solutions that the third problem admits. Firstly, the problem can be solved in two steps. One first uses some sort of normalization procedure to make the citations of articles in all fields at least approximately comparable. Then, one applies the citation indicator to each unit's normalized citation distribution. Secondly, consider the Top 10% indicator used in the construction of the influential Leiden and *SCImago* rankings. In the Leiden Ranking this indicator is defined as "*The proportion of publications of a university that, compared with other similar publications, belong to the top 10% most frequently cited…Publications are considered similar if they were published in the same field and the same publication and if they have the same document type*" (Waltman et al., 2012a). A similar definition is applied in the SCImago ranking (Bornmann et al., 2012) Note that this way of computing this particular indicator in the all-sciences case does not require any kind of prior citation normalization. For our purposes, it is useful to view this procedure as the average (weighted by the publication output) of the unit's Top 10% performance in each field. We note that this important precedent can be extended to *any* admissible indicator. Thus, given a classification system and an admissible citation indicator, we can compute the citation impact of a research unit in the all-sciences case as the appropriate weighted average of the unit's citation impact in each field. Independently of the conceptual interest of this proposal, we must compare the consequences of adopting it versus the possibility of following a normalization procedure.

Intuitively, the better the performance of the normalization procedure in eliminating the comparability difficulties across fields, the smaller will be the differences between the two approaches. Consider, for example, what we call the standard field-normalization procedure in which the normalized citations of articles in any field are equal to the articles' original raw citations divided by the field mean citation. Under the universality condition, that is, if field citation distributions were identical except for a scale factor, then the standard field-normalization procedure would completely eliminate all comparability difficulties. However, the universality condition, once claimed to be the case (Radicchi et al., 2008), is not usually satisfied in practice: even appropriately normalized, field citation distributions are seen to be significantly different from a statistical point of view (Albarrán et al., 2011a; and Waltman et al., 2012a). Therefore, at best, normalization procedures provide an approximate solution to the original comparability problem.

Using a measuring framework introduced in Crespo et al. (2013), recent research has established that different normalization procedures perform quite well in eliminating most of the effect in overall citation inequality that can be attributed to differences in production and citation practices between fields. This is the case for large Web of Science (WoS hereafter) datasets, classification systems at different aggregation levels, and different citation windows (Crespo et al., 2013, 2014; Li et al., 2013; Waltman & Van Eck, 2013; Ruiz-Castillo, 2014). The reason for the good performance of target (or cited-side) normalization procedures is that field citation distributions, although not universal, are extremely similar (Glänzel, 2007; Radicchi et al., 2008; Albarrán & Ruiz-Castillo, 2011; Albarrán et al., 2012; Waltman et al., 2012a; Radicci & Castellano, 2012; Li et al., 2013). It should be noted that this research on target normalization procedures uses WoS classification systems distinguishing at most between 235 sub-fields.

In principle, given the good performance of normalization procedures, we expect that the differences between the two approaches would be of a small order of magnitude.

Nevertheless, this is an empirical question that has never been investigated before. To confront this question, in this paper we conduct the following exercise.

- Ruiz-Castillo & Waltman (2015) apply the publication-level algorithmic methodology introduced by Waltman and Van Eck (2012) to a WoS hereafter dataset consisting of 9.4 million publications from the 2003-2012 period. This is done along 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 classification system recommended in Ruiz-Castillo and Waltman (2015), consisting of 5,119 clusters, of which 4,161 are referred to as significant clusters because they have more than 100 publications over this period. For the evaluation of research units' citation impact, we focus on the 3.6 million publications in the 2005-2008 period, and the citations they receive during a five-year citation window for each year in that period. It should be noted that, using the size- and scale-independent technique known as Characteristic Scores and Scales, Ruiz-Castillo and Waltman (2015) show that, as in previous research, significant clusters are highly skewed and similarly distributed.
- Our research units are the 500 universities in the 2013 edition of the CWTS Leiden Ranking (Waltman et al., 2012b). We analyze the approximately 2.4 million articles – about 67% of the total– for which at least one author belongs to one of these universities. We use a fractional counting approach to solve the problem –present in all classification systems– of the assignment of responsibility for publications with several co-authors working in different institutions. The total number of articles corresponding to the 500 universities is approximately 1.9 million articles –about 50% of the total.
- We evaluate the citation impact of each university using two size- and scale-independent indicators. Firstly, we use the Top 10% indicator, already mentioned. Secondly, one characteristic of this indicator is that it is not monotonic in the sense that it is invariant to any additional citation that a high-impact article might receive. Consequently, we believe that it is interesting to use a second indicator possessing this property. In particular, we select a member of the Foster, Greer, and Thorbecke (FGT hereafter) family, introduced in Albarrán et al. (2011b). We apply this indicator to the set of high-impact articles mentioned before. As will be seen below, the fact that both of our indicators are additively decomposable facilitates the comparability of the two solutions to the all-sciences aggregation problem.
- Using Crespo et al.'s (2013) measurement framework, Li et al. (2013) indicate that the best alternative among a wide set of target normalization procedures is the two-parameter system developed in Radicci and Castellano (2012). However, recent results indicate that the standard, one-parameter field-normalization procedure exhibits a good performance in reducing the effects on overall citation inequality attributed to differences in production and citation practices between fields (Radicchi et al., 2008; Crespo et al., 2013, 2014; Li et al., 2013; and Ruiz-Castillo, 2014). Consequently, in this paper we adopt this procedure in the usual solution to the all-sciences aggregation problem.
- We present two types of results. Firstly, we assess the performance of the standard normalization procedure in facilitating the comparability of the citations received by articles belonging to different clusters. Secondly, we assess the consequences of adopting the two solutions to the all-sciences aggregation problem by comparing the corresponding university rankings according to the two citation impact indicators.

The rest of the paper is organized into three sections. Section II presents the citation impact indicators, as well as the two solutions to the all-sciences aggregation problem. Section III describes the data, and includes the empirical results, while Section IV concludes.

The aggregation of heterogeneous scientific fields in the all-sciences case

Notation and citation indicators

It is convenient to introduce some notation. Given a set of articles S, and J scientific fields indexed by $j = 1,\ldots, J$, a *classification system* is an assignment of articles in S to the J fields. Let I be the number of research units, indexed by $i = 1,\ldots, I$. In this Section, the assignment of articles in S to the I research units is taken as given. Let $\boldsymbol{c_{ij}} = \{c_{ijk}\}$ be the *citation distribution of unit i in field j*, where c_{ijk} is the number of citations received by the k-th article, and let $\boldsymbol{c_j}$ be the *citation distribution of field j*, that is, the union of all research units' citation distributions in that field: $\boldsymbol{c_j} = \cup_i \{\boldsymbol{c_{ij}}\}$. Finally, let $\boldsymbol{C} = \cup_i \cup_j \{\boldsymbol{c_{ij}}\}$ be the *overall citation distribution*, or the citation distribution in the all-sciences case. For later reference, let N_{ij} be the number of articles in distribution $\boldsymbol{c_{ij}}$, let $N_i = S_j N_{ij}$ be the total number of articles published by unit i, let $N_j = S_i N_{ij}$ be the total number of articles in field j, and let $N = S_i S_j N_{ij}$ be the total number of articles in the all-sciences case.

A *citation impact indicator* is a function F defined in the set of all citation distributions, where $F(\boldsymbol{c})$ is the citation impact of distribution $\boldsymbol{c}$. Let $\boldsymbol{c^r}$ be the r-th replica of distribution $\boldsymbol{c}$. An indicator F is said to be *size-independent* if, for any citation distribution $\boldsymbol{c}$, $F(\boldsymbol{c^r}) = F(\boldsymbol{c})$ for all r. An indicator F is said to be *scale-independent* if for any $\lambda > 0$, and any citation distribution $\boldsymbol{c}$, $F(\lambda\boldsymbol{c}) = F(\boldsymbol{c})$. An indicator F is said to be *additively decomposable* if for any partition of a citation distribution $\boldsymbol{c}$ into G sub-groups, indexed by $g = 1,..., G$, the citation impact of distribution $\boldsymbol{c}$ can be expressed as follows:

$$F(\boldsymbol{c}) = S_g\, (M_g/M) F(\boldsymbol{c_g}),$$

where M_g is the number of publications in sub-group g, and $M = \Sigma_g\, M$ is the number of publications in distribution $\boldsymbol{c}$.

Consider the following two difficulties for comparing the citation impact of any pair of research units: the two units may be of different sizes, and if they work in different fields, then their raw citations are not directly comparable. As it is well known, these two difficulties can be overcome using a size- and scale-independent indicator. The following two indicators are good examples of size- and scale-independent indicators that, in addition, are additively decomposable.

1. Let $\boldsymbol{X_j}$ be the set of the 10% most cited articles in citation distribution $\boldsymbol{c_j}$, and let $\boldsymbol{x_{ij}}$ be the sub-set of articles in $\boldsymbol{X_j}$ corresponding to unit i, so that $\boldsymbol{X_j} = \cup_i\{\boldsymbol{x_{ij}}\}$ with $\boldsymbol{x_{ij}}$ non-empty for some i. If n_{ij} is the number of articles in $\boldsymbol{x_{ij}}$, then the *Top 10% indicator for unit i in field j*, T_{ij}, is defined as

$$T_{ij} = n_{ij}/N_{ij}. \qquad (1)$$

Of course, for field j as a whole, if $n_j = \Sigma_i\, n_{ij}$ is the number of articles in $\boldsymbol{X_j}$, then $T_j = n_j/N_j = 0.10$.

2. Let z_j be the *Critical Citation Line* –CCL hereafter– for citation distribution $\boldsymbol{c_j}$, and denote the articles in $\boldsymbol{c_j}$ with citations equal to or greater than z_j as *high-impact articles*. For any high impact article with citations c_{il}, define the *CCL normalized high-impact gap* as $(c_{il} - z_j)/z_j$.

Consider the family of FGT indicators introduced in Albarrán et al. (2011b) as functions of normalized high-impact gaps. The second member of this family, A_{ij}, is defined as

$$A_{ij} = (1/N_{ij})[S_l(c_{il} - z_j)/z_j], \quad (2)$$

where the sum is over the high-impact articles in citation distribution $\boldsymbol{c_j}$ that belong to unit i. We refer to this indicator as the *Average of high-impact gaps for unit i in field j*. For the entire field j as a whole, the average of high-impact gaps is defined as

$$A_j = (1/N_j)[S_k (c_k - z_j)/z_j],$$

where the sum is over the high-impact articles in citation distribution $\boldsymbol{c_j}$.

To facilitate the comparison with T_{ij}, in the sequel we will always fix z_j as the number of citations of the article in the 90th percentile of citation distribution $\boldsymbol{c_j}$. In that case, the set of high-impact articles coincides with the set of the 10% most cited articles in citation distribution $\boldsymbol{c_j}$. The two main differences between the two indicators are the following. Firstly, one or more citations received by a high-impact article increases A_{ij} but does not change T_{ij}. In other words, A_{ij} is monotonic but T_{ij} is not. Secondly, T_{ij} is more robust to extreme observations than A_{ij}.

The solution to the all-sciences aggregation problem using the standard field-normalization procedure

For any i, let $\boldsymbol{c_i} = (\boldsymbol{c_{i1}},\ldots, \boldsymbol{c_{ij}},\ldots, \boldsymbol{c_{iJ}})$ be the *raw citation distribution of unit i in the all-sciences case*. Differences in production and citation practices across fields make impossible the direct comparison of the raw citations received by articles in different fields. In order to achieve some comparability, one possibility is to use some normalization procedure. For any article k in citation distribution $\boldsymbol{c_{ij}}$, the normalized number of citations c^*_{ijk} according to the *standard field-normalization procedure* is defined as

$$c^*_{ijk} = c_{ijk}/\mu_j.$$

For any i, let $\boldsymbol{c^*_i} = \cup_j \cup_k \{c^*_{ijk}\} = (\boldsymbol{c^*_{i1}},\ldots, \boldsymbol{c^*_{ij}},\ldots, \boldsymbol{c^*_{iJ}})$ be the *normalized citation distribution of unit i in the all-sciences case*. Since normalized citations are now comparable, it makes sense to apply any indicator to citation distribution $\boldsymbol{c^*_i}$. For any i, let $\boldsymbol{F^*_i} = F(\boldsymbol{c^*_i})$ be the citation impact of distribution $\boldsymbol{c^*_i}$ according to the indicator F. For any pair of research units u and v in the all-sciences case, the citation impact values $\boldsymbol{F^*_u}$ and $\boldsymbol{F^*_v}$ are now comparable, and can be used to rank the two units in question.

Note that, since $\boldsymbol{c^*_i}$ for $i = 1,\ldots, I$ forms a partition of $\boldsymbol{C^*}$ and F is assumed to be additively decomposable, we can write

$$F^* = F(\boldsymbol{C^*}) = S_i (N_i/N)\boldsymbol{F^*_i}.$$

Thus, if we rank universities by the ratio $\boldsymbol{F^*_i}/F^*$, $i = 1,\ldots, I$, then the value one can serve as a benchmark for evaluating the research units in the usual way. For later reference, since $\boldsymbol{c^*_{ij}}$ for $j = 1,\ldots, J$ forms a partition of $\boldsymbol{c^*_i}$, for each i we can write

$$\boldsymbol{F^*_i} = F(\boldsymbol{c^*_i}) = S_j (N_{ij}/N_i)\boldsymbol{F^*_{ij}}, \quad (3)$$

where $F^*_{ij} = F(\boldsymbol{c}^*_{ij})$ for all j, that is, F^*_{ij} is simply the citation impact of citation distribution $\boldsymbol{c}^*_{ij}$ according to F.

A solution to the all-sciences aggregation problem without field-normalization

For any i and any j, denote by $F_{ij} = F(\boldsymbol{c}_{ij})$ the citation impact of distribution $\boldsymbol{c}_{ij}$ according to F. A convenient measure of citation impact for unit i in the all-sciences case, F_i, can be defined as the weighted average of the values F_{ij} achieved in all fields, with weights equal to the relative importance of each field in the total production of unit i:

$$F_i = S_j\,(N_{ij}/N_i)F_{ij} \qquad (4)$$

The comparison of expressions (4) and (5) illustrate the differences between the two solutions to the all-sciences aggregation problem when the evaluation of the units' citation impact is made with additively decomposable indicators. Finally, it is convenient to compute the weighted average of these quantities as follows:

$$F = S_i\,(N_i/N)F_i.$$

Thus, as before, if we rank universities by the ratio F_i/F, $i = 1,\ldots, I$, then the value one can serve as a benchmark for evaluating the research units in the usual way. In practice, we have information concerning some but not all research units. Therefore, we compute F as the following weighted average: $F = \mathrm{S}_j\,(N_j/N)F_j$, where $F_j = \mathrm{F}(\boldsymbol{c}_j)$.

The aim of the paper

The main aim of this paper is the comparison between the rankings of research units obtained with and without the standard field-normalization procedure, $(F^*_1, \ldots, F^*_I)$ and $(F_1, \ldots, F_I)$, respectively.

To understand the way the results will be presented, we need to review the connection between the performance of the normalization procedure and the relationship between the solutions to the all-sciences aggregation problem. For that purpose, we need to introduce some more notation. For any j, let $\boldsymbol{x}_j$ be the set of high-impact articles in distribution $\boldsymbol{c}_j$, that is, the set of articles in $\boldsymbol{c}_j$ with citations equal to or greater than z_j, or the set of the 10% most cited articles in $\boldsymbol{c}_j$. Let us denote by $\boldsymbol{X} = (\boldsymbol{x}_1,\ldots, \boldsymbol{x}_j,\ldots, \boldsymbol{x}_J)$ the set of high-impact articles in the all-sciences case. On the other hand, let $\boldsymbol{Y}$ be the set of the 10% most cited articles in the overall normalized citation distribution $\boldsymbol{C}^* = \cup_j \{\boldsymbol{c}^*_j\}$. Let $\boldsymbol{y}_j$ be the sub-set of articles in $\boldsymbol{Y}$ belonging to field j, so that $\boldsymbol{Y} = (\boldsymbol{y}_1,\ldots, \boldsymbol{y}_j,\ldots, \boldsymbol{y}_J)$. Note that, in practice, the sets $\boldsymbol{y}_j$ might be empty for some j.

Under the universality condition, that is, if all fields are equally distributed except for a scale factor then, at every percentile of field citation distributions, normalized citations will be the same for all fields. In other words, the normalization procedure will work perfectly. In particular, in this situation we would have $z_j/\mu_j = z^*$ for all j. Consequently, we would have $\boldsymbol{y}_j = \boldsymbol{x}_j$ for all j, and $\boldsymbol{Y} = \boldsymbol{X}$. Since citation distributions $\boldsymbol{c}^*_{ij}$ and $\boldsymbol{c}_{ij}$ have the same number of articles and our indicators are a function solely of high-impact articles, we would have $F^*_{ij} = F(\boldsymbol{c}^*_{ij}) = F_{ij} = F(\boldsymbol{c}_{ij})$ for all i and j. In view of equations (4) and (5), we would have $F^*_i = F_i$ for all i. In other words, the rankings $(F^*_1, \ldots, F^*_I)$ and $(F_1, \ldots, F_I)$ will be identical.

As we know, in practice the universality condition is not satisfied. However, the better the performance of the normalization procedure, that is, the closer is the set $\boldsymbol{Y}$ to set $\boldsymbol{X}$, the more similar the rankings $(F^*_1, \ldots, F^*_I)$ and $(F_1, \ldots, F_I)$ are expected to be for any F. Note that this

conjecture has to be verified in practice. In any case, the empirical section begins by assessing the performance of the normalization procedure.

On the other hand, independently of the normalization procedure's performance, we should measure the consequences of adopting the two solutions to the all-sciences aggregation problem using indicators with different properties. The reason, of course is that whenever $\boldsymbol{Y}$ and $\boldsymbol{X}$ differ, that is, when the set of high-impact articles under the two solutions differ, the consequences for the university rankings might be of a different order of magnitude depending on the citation impact indicator we use. This is the reason why we will study the situation using the Top 10% and the Average of high-impact gaps.

Empirical results

The data and descriptive statistics

As indicated in the Introduction, our dataset results from the application of a publication-level methodology to 9,446,622 distinct articles published in 2003-2012 (see Ruiz-Castillo & Waltman, 2015). Publications in local journals, as well as popular magazines and trade journals have been excluded (see Ruiz-Castillo & Waltman, 2015 for details). 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. The classification system consists of 5,119 clusters, and citation distributions refer to the citations received by these articles during a five-year citation window for each year in that period. In this paper, we focus on the set of 3,614,447 distinct articles published in 2005-2008. In terms of the notation introduced in Section II.1, we have $\boldsymbol{C} = \cup_j \{\boldsymbol{c_j}\} = (c_1,\ldots, c_N)$ with J = 5,119, and N = 3,614,447.

The research units are universities. Publications are assigned to universities using the fractional counting method that takes into account the address lines appearing in each publication. An article is fully assigned to a university only if all addresses mentioned in the publication belong to the university in question. If a publication is co-authored by two or more universities, then it is assigned fractionally to all of them in proportion to the number of address lines. For example, if the address list of an article contains five addresses and two of them belong to a particular university, then 0.4 of the article is assigned to this university, and only 0.2 of the article is assigned to each of the other three universities.

We know the total number of address lines of every publication, but we have information about the number of address lines of specific institutions only for the 500 LR universities. This number is well below I, the total number of research units in the notation introduced in Section II.1. There are 2,420,054 distinct articles, or 67% of the total, with at least one address line belonging to a LR university. The total number of articles in the LR universities according to the fractional counting method is 1,886,106.1, or 52.2% of the total. The distribution of this total among the 500 universities is available in Perianes-Rodriguez & Ruiz-Castillo, 2014a.

The performance of the normalization procedure

We assess the performance of the normalization procedure using the measurement framework introduced in Crespo et al. (2013), we first estimate the effect on overall citation inequality attributable to differences in production and citation practices between clusters, and then the reduction in this effect after applying the standard field-normalization procedure. Given the many clusters with very few publications (see Ruiz-Castillo & Waltman, 2015), we apply this method to the 3,332 clusters with more than 250 publications. These clusters include 3,441,666 million publications, or 95.2% of the total.

We begin with the partition of, say, each cluster citation distribution into P quantiles, indexed by $p = 1,.., P$. In practice, in this paper we use the partition into percentiles, that is, we choose $P = 100$. Assume for a moment that, in any cluster i, we disregard the citation inequality within every percentile by assigning to every article in that percentile the mean citation of the percentile itself, μ_i^p. The interpretation of the fact that, for example, $\mu_i^p = 2\ \mu_j^p$ is that, on average, the citation impact of cluster i is twice as large as the citation impact of cluster j in spite of the fact that both quantities represent a common underlying phenomenon, namely, the same *degree of citation impact* in both clusters. In other words, for any π, the distance between μ^p and μ_j^p is entirely attributable to the difference in the production and citation practices that prevail in the two clusters for publications with the same degree of excellence in each of them. Thus, the citation inequality between clusters at each percentile, denoted by $I(p)$, is entirely attributable to the differences in citation practices between the 3,332 clusters holding constant the degree of excellence in all clusters at quantile π. Hence, any weighted average of these quantities, denoted by *IDCC* (*I*nequality due to *D*ifferences in *C*itation impact between *C*lusters), provides a good measure of the total impact on overall citation inequality that can be attributed to such differences. Let $\boldsymbol{C'}$ be the union of the clusters citation distributions, $\boldsymbol{C'} = \cup \{\boldsymbol{c_j}\}$ for $j = 1,\ldots, 3{,}332$. We use the ratio

$$IDCC/I(\boldsymbol{C'}) \qquad (6)$$

to assess the relative effect on overall citation inequality, $I(\boldsymbol{C'})$, attributed to the differences in citation practices between clusters (for details, see Crespo et al., 2013).

Finally, we are interested in estimating how important scale differences between cluster citation distributions are in accounting for the effect measured by expression (6). For that purpose, we use the relative change in the *IDPC* term, that is, the ratio

$$[IDCC - IDCC^*]/IDCC, \qquad (7)$$

where *IDCC** is the term that measures the effect on overall citation inequality attributed to the differences in cluster distributions after applying the standard field-normalization procedure (for details, see again Crespo et al., 2013). The estimates of expressions (6) and (7) are as follows:

Table 1. The effect on overall citation inequality, I('C), of the differences in citation impact between clusters before and after standard field-normalization, and the impact of normalization on this effect.

		Normalization impact =100 [*IDCC – IDCC*/IDCC*]
Before MNCS normalization, 100 [*IDCC/I(C')*]	22.5 %	-
After MNCS normalization, 100 [*IDCC*/I(C')*]	4.3 %	84.3 %

It can be observed that the effect of the differences in citation practices between such a large number of clusters represents 22.5% of overall citation inequality, a figure much larger than what has been found in the previous literature for at most 235 sub-fields. Nevertheless, the standard field-normalization procedure reduces this effect down to 4.3%, quite an achievement.

Differences in university rankings under the two solutions to all-sciences aggregation problem

The university rankings without and with normalization according to the Top 10% indicator, T_i and T^*_i, and according to the Average of high-impact gaps, A_i and A^*_i can be found in Perianes-Rodriguez & Ruiz-Castillo (2014a). We begin with the comparison of university rankings according to T_i and T^*_i. The Pearson correlation coefficient between university values is 0.995, while the Spearman correlation coefficient between ranks is 0.992. However, high correlations between university values and ranks do not preclude important differences for individual universities. In analyzing the consequences of going from T_i to T^*_i, we must take two aspects into account. Firstly, we should analyze the re-rankings that take place in such a move. Secondly, we should compare the differences between the university values themselves. Fortunately, we have a relevant instance with which to compare our results: the differences found in Ruiz-Castillo and Waltman (2015) in going from the university rankings according to T_i using the Web of Science classification system with 236 journal subject categories, or sub-fields, and the classification system we are using in this paper with 5,119 clusters.

As much as 38.4% of universities experience very small re-rankings of less than or equal to five positions, while 67 universities, or 13.4% of the total, experience re-rankings greater than 25 positions. These figures are 20.2% and 39.0% when going from the WoS classification system to our dataset. Among the first 100 universities, 61 experience small re-rankings in going from T_i to T^*_i, while only 44 are in this situation in the change between classification systems. As far as the cardinal changes is concerned, 78.4% of universities have changes in top 10% indicator values smaller than or equal to 0.05 when going from T_i to T^*_i. This percentage is 71% among the first 100 universities. These figures are 50.1% and 60.0% in the change between classification systems. For most universities, the differences are more or less negligible. Although for some universities more significant differences can be observed, the conclusion is clear. The differences observed in university rankings according to the top 10% indicator when we adopt the two solutions for solving the all-sciences aggregation problem are considerably less than according to the same indicator when we move from the WoS classification system to our dataset (Perianes-Rodriguez & Ruiz-Castillo, 2014a).

The Pearson correlation coefficient between the university rankings according to the average of high-impact gaps, A_i and A^*_i, is 0.596, while the Spearman correlation coefficient between ranks is 0.984. However, the low Pearson correlation coefficient is due to the presence of the well-known extreme observation of the University of Göttingen (Waltman et al., 2012b; Ruiz-Castillo & Waltman, 2015). Without this university, this correlation coefficient becomes 0.986. In any case, as before, high correlations between university values and ranks do not preclude important differences for individual universities. The ordinal differences in university rankings according to this indicator with and without field-normalization are of a similar order of magnitude as those obtained with the top 10% indicator. For example, 33.0% of universities experience very small re-rankings of less than or equal to five positions, while 80 universities, or 16.0% of the total, experience re-rankings greater than 25 positions. Among the first 100 universities, only 44 experience small re-rankings in going from A_i to A^*_i (in comparison with 61 when going from T_i to T^*_i). As far as the cardinal changes is concerned, 64.2% of universities have changes in indicator values smaller than or equal to 0.05 when going from A_i to A^*_i –a comparable figure with 78.4% when going from T_i to T^*_I (Perianes-Rodriguez & Ruiz-Castillo, 2014a).

The conclusion is inescapable. In spite of the fact of the limitations of the standard normalization procedure in the presence of so many clusters, the differences observed in university rankings when we adopt the two solutions for solving the all-sciences aggregation problem are of a relatively small order of magnitude regardless of which of then two rather different citation impact indicators is used in obtaining the university rankings.

Conclusions

The heterogeneity of the fields distinguished in any classification system poses a severe aggregation problem when one is interested in evaluating the citation impact of a set of research units in the all-sciences case. In this paper, we have analyzed two possible solutions to this problem. The first solution relies on prior normalization of the raw citations received by all publications. In particular, we focus on the standard field-normalization procedure in which field mean citations are used as normalization factors. The second solution extends the approach adopted in the Leiden and SCImago rankings for computing the Top 10% indicator in the all-sciences case to any admissible indicator. This solution does not require any prior field-normalization: the citation impact of any research unit in the all-sciences case is calculated as the appropriately weighted sum of the citation impact that the unit achieves in each field.

Using a large WoS dataset consisting of 3.6 million publications in the 2005-2008 period and an algorithmically constructed publication-level classification system that distinguishes between 5,119 clusters, this simple alternative has been confronted with the usual one when the citation impact of the 500 LR universities are evaluated using two indicators with very different properties: the top 10% indicator, and the average of high-impact gaps.

The shape of the citation distributions of 4,161 significant clusters with more than 100 publications in our dataset has been previously shown to be highly skewed and reasonable similar (Ruiz-Castillo & Waltman, 2015). Previous results with WoS classification systems that distinguishes at most between 235 sub-fields indicate that, when this is the case, the standard field-normalization procedure performs well in reducing the overall citation inequality attributed to the differences in production and citation practices between fields. In this paper we have shown that this is not exactly the case, even when we restrict the attention to 3,332 clusters with more than 250 publications. Therefore, a priori it was not obvious what to expect when confronting the solutions to the all-sciences aggregation problem with and without prior field-normalization.

Interestingly enough, the differences between the university rankings obtained with both solutions is of a relatively small order of magnitude independently of the citation impact indicator used in the construction of the university rankings. In particular, these differences are considerably smaller than the ones obtained in Ruiz-Castillo and Waltman (2015) for the move from the WoS classification system with 236 sub-fields to the one used in this paper with 5,119 clusters.

In principle, it seems preferable to evaluate the citation impact of research units in the all-sciences case avoiding any kind of prior normalization operation. However, the empirical evidence presented in this paper indicates that that the use of the traditional methodology does not lead to very different results. This is a convenient conclusion, since there are instances when normalization is strongly advisable. For example, when one is interested in studying the research units citation distributions in the all-sciences case –as we do in the companion paper Perianes-Rodriguez and Ruiz-Castillo (2014b).

It should be noted that, before being accepted, it would be convenient to replicate the results of this paper for other datasets, other classification systems, other types of research units, and other ways of assigning responsibility between research units in the case of co-authored publications.

References

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. (2011a). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, *88*, 385–397.

Albarrán, P., Ortuño, I., & Ruiz-Castillo, J. (2011b). The measurement of low- and high-impact in citation distributions: technical results. *Journal of Informetrics*, *5*, 48–63.

Bornmann, L., De Moya Anegón, F., & Leydesdorff, L. (2012) The new excellence indicator in the world report of the SCImago Institutions Rankings 2011. *Journal of Informetrics*, *6*, 333-335.

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 Association 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.

Li, Y., Castellano, C., Radicchi, F., & Ruiz-Castillo, J. (2013). Quantitative evaluation of alternative field normalization procedures. *Journal of Informetrics*, *7*, 746–755.

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2014a). *An alternative to field-normalization in the aggregation of heterogeneous scientific fields*. Working Paper, Economic Series 14-25, Universidad Carlos III (http://hdl.handle.net/10016/19812).

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2014b). *University citation distributions*. Working Paper, Economic Series 14-26, Universidad Carlos III (http://hdl.handle.net/10016/19811).

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", *PNAS*, *105*, 17268-17272.

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., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, *9*, 102-117. (DOI: 10.1016/j.joi.2014.11.010).

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., & Van Raan, A. F. J. (2012a). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, *63*, 72–77.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., Van Eck, N. J., Van Leeuwen, T. N., Van Raan, A. F. J., Visser, M. S., & Wouters, P. (2012b). 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. (2013). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, *7*, 833–849.