Journal of Informetrics 8 (2014) 917–934

Contents lists available at ScienceDirect

Journal of Informetrics

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

The skewness of scientific productivity

Javier Ruiz-Castillo [a,*], Rodrigo Costas [b]

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

ARTICLE INFO

Article history:
Received 8 May 2014
Received in revised form 15 August 2014
Accepted 22 September 2014
Available online 14 October 2014

Keywords:
Individual scientist's productivity distributions
Skewness of science
Disambiguation algorithm
Co-authorship

ABSTRACT

This paper exploits a unique 2003–2011 large dataset, indexed by Thomson Reuters, consisting of 17.2 million disambiguated authors classified into 30 broad scientific fields, as well as the 48.2 million articles resulting from a multiplying strategy in which any article co-authored by two or more persons is wholly assigned as many times as necessary to each of them. The dataset is characterized by a large proportion of authors who have their *oeuvre* in several fields. We measure individual productivity in two ways that are uncorrelated: as the number of articles per person and as the mean citation per article per person in the 2003–2011 period. We analyze the shape of the two types of individual productivity distributions in each field using size- and scale-independent indicators. To assess the skewness of productivity distributions we use a robust index of skewness, as well as the Characteristic Scores and Scales approach. For productivity inequality, we use the coefficient of variation. In each field, we study two samples: the entire population, and what we call "successful authors", namely, the subset of scientists whose productivity is above their field average. The main result is that, in spite of wide differences in production and citation practices across fields, the shape of field productivity distributions is very similar across fields. The parallelism of the results for the population as a whole and for the subset of successful authors, when productivity is measured as mean citation per article per person, reveals the fractal nature of the skewness of scientific productivity in this case. These results are essentially maintained when any article co-authored by two or more persons is fractionally assigned to each of them.

1. Introduction

In this paper, we study the size and the mean of individual citation distributions in a given period of time for all authors in a number of scientific fields. Naturally, the size of individual citation distributions, that is, the number of publications per author, is a standard measure of individual productivity. The productivity of individual scientists has been studied extensively since Lotka's (1926) pioneer contribution, in which the probability of an author publishing a certain number of articles in Chemistry was estimated to be an inverse square function of the number of publications (Alvarado, 2012, counts 651 publications concerning the so-called *Lotka's law* from that date until 2010). However, most of these contributions

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

http://dx.doi.org/10.1016/j.joi.2014.09.006
1751-1577/

analyze a relatively small number of scientists and, to the best of our knowledge, do not systematically study productivity distributions using comparable and large datasets for several scientific disciplines.[1]

On the other hand, the mean citation per article per author is a standard (size-independent) measure of the citation impact achieved by any researcher in her field of study. Nevertheless, to simplify the exposition, we will refer to this indicator of citation impact as a second definition of individual productivity. At any rate, we do not know of systematic studies concerning the distribution of this variable within and between representative samples for a variety of scientific disciplines.

As in any other scientific discipline, in Scientometrics we should clearly establish the stylized facts that characterize basic constructs in all fields. Consequently, this paper studies the productivity of individual scientists – in the two senses indicated above – in 30 broad fields using a large dataset, indexed by Thomson Reuters, consisting of 7.7 million distinct articles published in the period 2003–2011 in academic journals. Applying a variable citation window from the publication year until 2012, these articles receive 78.9 million citations.

Regardless of how we measure individual productivity, a study of this type must confront the following four methodological problems: (i) the classification of articles into scientific fields; (ii) the identification of the author(s) of each article, (iii) the allocation of authors to fields, and (iv) the attribution of individual responsibility in cases of multiple authorship. After these problems are solved (see Section 2), we end up with a dataset consisting of 17.2 million authors and 48.2 million articles.

Of course, we know a priori that the between-field variability with respect to several basic characteristics is typically very large. Firstly, the size of productivity distributions, namely, the number of authors per field, is bound to be very different across fields. Secondly, because of well-known differences in production and citation practices, the average number of articles per author, as well as the average mean citation per article per author are also expected to be very different across fields.

Therefore, what we should study is the shape of field productivity distributions abstracting from size and scale differences across fields. To simplify the presentation, we focus on the skewness of productivity distributions. Naturally, the extensive literature on Lotka's law leads us to expect that productivity distributions according to the first definition are highly skewed in all fields, in the sense that a majority of individuals publish very little, while a large proportion of the total number of publications must be attributed to a small number of authors. Finally, if only by analogy with the skewness of science in so many dimensions (see De Solla Price, 1963; Lotka, 1926; Seglen, 1992, to cite only a few classics), we expect that all field productivity distributions according to the second definition are also highly skewed.

In this scenario, the main aim of this paper is to investigate the between-field variation of the skewness of productivity distributions that is expected to be prevalent in each field. For the reasons already explained, we need size- and scale-independent indicators of skewness. We follow two complementary approaches. In the first place, we study the broad features of this phenomenon by simply partitioning productivity distributions into three classes of individuals with low, fair, and very high productivity. For this purpose, we adopt the Characteristic Scores and Scale (CSS hereafter) approach first introduced in Scientometrics by Schubert, Glänzel, and Braun (1987). In the second place, we are interested in summarizing the skewness of productivity distributions with a single scalar. Among the size- and scale-independent skewness measures that are also robust to extreme observations, in this paper we use the one suggested by Groeneveld and Meeden (1984) that has been used before in Albarrán, Perianes-Rodriguez, and Ruiz-Castillo (2014), and Perianes-Rodriguez and Ruiz-Castillo (2014). Finally, for reasons that will be apparent in the sequel, we analyze the shape of productivity distributions in each field for two samples: the entire population, and what we call *successful authors*, namely, the subset of scientists whose productivity is above their field average.

In the Working Paper version of this article, Ruiz-Castillo and Costas (2014), hereafter referred to as RCC, we study a second characteristic of the shape of field productivity distributions: the productivity inequality exhibited both by the entire population and successful authors for the two productivity definitions. A summary of results is presented below in a section on extensions. Also, to facilitate the reading of the text, some statistical information and, in many cases, the numerical results for a variety of field characteristics, are relegated to the Supplementary Material Section (SMS hereafter) of the paper. At the end of each section we include a footnote specifying which aspects of the questions discussed in the text can be found in the SMS.

The rest of this paper is organized into five sections. Section 2 describes the data and discusses our approaches to cope with the four methodological issues. Sections 3 and 4 present the results concerning the characteristics of productivity distributions when individual productivity is measured as the number of publications and as the mean citations per article, respectively, whereas Section 5 summarizes the main results concerning productivity inequality and other issues explored in detail in RCC. Finally, Section 6 summarizes the paper and suggests possible extensions.

[1] Kyvic (1989) compares the productivity between three very broad scientific disciplines – the Medical, the Natural, and the Social Sciences – and the Humanities, using a relatively small dataset. A key exemption is the important contribution by Ioannidis et al. (2014), which studies 15.1 million authors that have published at least one indexed item in the entire Scopus database in the period 1996–2011. See below for a comparison of our methods and results with those of Ioannidis et al. (2014).

2. Methodological issues

Since we wish to address a homogeneous population, in this paper only research articles published in academic journals or, simply, *articles* are studied.[2] As indicated in Section 1, we begin with a large sample, consisting of 7,721,132 distinct articles published in the period 2003–2011. In what follows, we discuss the solutions we have adopted for coping with the four methodological problems mentioned in Section 1.

1. Given the well-known differences in publication and citation practices across scientific disciplines, the performance of any pair of authors can only be compared if they belong to the same field. The problem, of course, is that Thomson Reuters assigns publications in the periodical literature to Web of Science subject categories via the journal in which they have been published. Many journals are assigned to a single category, but many others are assigned to two, three, or even more categories up to a maximum of six. In particular, in our dataset 2,246,435 articles, or 29.1% of the total, are assigned to two or more of our 30 fields.

   There are two approaches to tackle the problem created by the assignment of publications to two or more subject categories in Thomson Reuters datasets. The first is a fractional strategy, where each publication is fractioned into as many equal pieces as necessary with each piece assigned to its corresponding sub-field. The second approach follows a multiplicative strategy in which each paper is counted as many times as necessary in the several sub-fields to which it is assigned. In this way, the space of articles is expanded as much as necessary beyond the initial size in what we call the *extended count*. Fortunately, previous results indicate that for many purposes journals assigned to a single or to several subject categories share similar characteristics, so that the strategy choice is not that crucial. Among other issues in citation analysis, the study of the skewness of citation distributions across fields at different aggregation levels, or the evaluation of the gap in citation impact between the U.S. and the European Union using different indicators, are very robust to the strategy selected (Crespo, Herranz, Li, & Ruiz-Castillo, 2014; Herranz & Ruiz-Castillo, 2012a, 2012b, 2012c, 2013). All in all, in this paper we follow a multiplicative strategy. Consequently, the number of articles in the corresponding extended count is 10,355,901, or 34.1% larger than the number of distinct articles (cf. Table I in SMS).

   On the other hand, it is well known that there is no generally agreed-upon Map of Science or aggregation scheme that allows us to ascend from Web of Science categories up to other aggregate levels. Among the many alternatives, we take as our starting point the partition of scientific activity into the 35 broad fields introduced by Tijssen, Hollanders, and van Steen (2010) that has been used in Buter and van Raan (2011), Hoekman, Frenken, and Tijssen (2010), Hoekman, Scherngell, Frenken, and Tijssen (2013), and Schneider and Costas (2013). We exclude the following five fields from this list because of their limited coverage in the Web of Science database used in this paper: Creative Arts, Culture & Music; History, Philosophy & Religion; Language & Linguistics; Literature; Political Science & Public Administration. Therefore, we end up with the remaining 30 fields.[3]
2. The accurate assignment of articles to individual authors is known to be plagued with formidable obstacles (Costas, Van Leeuwen, & Bordons, 2010; Lindsey, 1980). In this paper, we solve this problem using the algorithm recently generated by Caron and van Eck (2014). This is an author disambiguation algorithm for large bibliometric databases that belongs to what is known in the literature as the class of unsupervised learning approaches. The method, inspired in Levin, Krawczyk, Bethard, and Jurafsky (2012), uses rule-based scoring and clustering of the individuals' publications, thus detecting their *oeuvres* in a systematic and accurate way. Although the clustering is not perfect, we believe that we are working with quite realistic data concerning the assignment of articles to individual scholars.[4] Overall, there are 9,631,769 distinct researchers associated to the 7.7 million distinct articles of the dataset.
3. Given the allocation of articles to fields in a multiplicative way, in this paper the authors of each article are allocated to fields in the same multiplicative way. Therefore, in the extended count the number of authors goes up to 17,199,433 individuals, a 78.6% increase relative to the original number of distinct authors. In order to facilitate the interpretation of our results, it is important to clarify the consequences of this procedure for the extent in which researchers appear as authors in several fields.[5]

   Consider the 5,474,693 articles, or 70.9% of the total, that are assigned to a single field in our dataset (cf. Table I in SMS). These articles are written by 7,555,176 distinct authors, or 78.4% of the total. However, some of these researchers are authors of articles that belong to different fields. Therefore, the total number of authors assigned to the 30 fields is somewhat larger: 9,472,725. In our view, this poses no problem of interpretation: for the purpose of analyzing the characteristics of authors in a number of different fields, as we do in this paper, researchers who write articles in several fields should be treated as independent, different authors in their respective fields.

[2] Following Waltman and Van Eck (2013a, 2013b), we exclude publications in local journals, as well as magazine and trade journals.

[3] It is not claimed that this scheme provides the best possible representation of the structure of science. It is rather a convenient simplification for the discussion of field comparability issues in this paper.

[4] The clustering method is conservative in the sense that it values precision over recall. Specifically, the evaluation of the method shows on average a 95% precision and a 90% recall (cf. Caron & van Eck, 2014).

[5] We thank one referee for pointing to us out the importance of this issue.

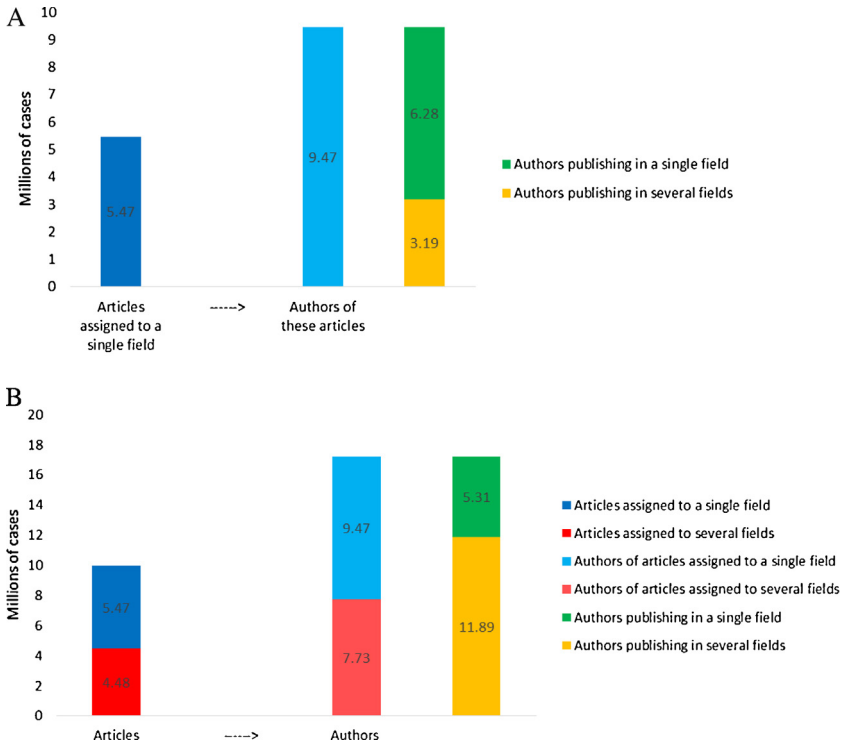

Fig. 1. (A) Authors (of articles assigned to a single field) who have their *oeuvre* in one or more fields. (B) Authors (and articles in the extended count) who have their *oeuvre* in one or more fields.

Consider now the 2,246,439 distinct articles assigned to two o more of the 30 fields. In the multiplicative approach, this subset gives rise to 4,881,208 extended articles. The total number of scholars writing them in the different fields is 7,726,708. Together with the 9.5 million authors introduced in the previous paragraph, this gives rise to the 17.2 million authors in the extended count (cf. Table IV in SMS, column 3). Naturally, by construction, these 7.7 million authors appear in two or more fields – a situation that would directly increase the proportion of authors whose *oeuvre* appears in several fields. Moreover, some of these 7.7 million researchers would be part of the 9.5 million already studied. Consequently, some of the scholars who had all their publications in one field would now have some articles in several fields. The end result is that only 5,306,383 authors in the extended count, or 30.8% of the total, have all their *oeuvre* in a single field. The situation – which we have not seen discussed before in the literature – is illustrated in Fig. 1.

It could be argued that, given any classification system that distinguishes between a minimum number of scientific fields, there are *some* publications that must be simultaneously classified into several fields. Although we have no means of knowing the true extent of this phenomenon, the percentage of authors in several fields in a classification system where each article is assigned to a single field would give us a lower bound for the true situation. Instead, in the case of our Thomson Reuters dataset, where an important percentage of articles are assigned to two or more fields, it is likely that the percentage of authors whose *oeuvre* appears in several fields – which is equal to 69.2% of the total – exaggerates the true situation.[6]

4. A fundamental difficulty in the study of scientists' productivity is the definition of the individual contribution to an article in a world dominated by multiple authorship in all fields (Cronin, 2001). The following two points should be noted. Firstly, the mean number of authors per article ranges from 1.7 or 1.9 in seven fields to 4.5 or 4.8 in five fields (cf. Table III in SMS), with a maximum of 5.3 in Astronomy & Astrophysics (cf. Table III in SMS). The average over the 30 fields is 3.1 with a relatively high coefficient of variation of 0.35. Secondly, the maximum number of authors per article reveals several truly extreme observations: it is greater than 3000 in Physics & Materials Sciences, Multidisciplinary Journals, and Astronomy & Astrophysics, and greater than 2450 in Instruments & Instrumentation, and Clinical Medicine. At the opposite end, the maximum number of authors in Mathematics is 36, while in General & Industrial Engineering, and Information & Communication Sciences it is 26.

[6] Note that the large percentage of authors with publications in several fields in our dataset is independent of the multiplicative or fractional approach one adopts to articles assigned by Thomson Reuters to several subject categories.

In this situation, arbitrarily choosing a single author per article without even the assurance that s/he is the "leading author" is out of the question. On the other hand, an adjusted or fractional count introduces measurement on a continuous scale, perhaps inappropriate for a phenomenon that is clearly discrete, and perhaps even representing a scale with a degree of precision greater than we are actually capable of measuring (Nicholls, 1989). Finally, as shown by Rousseau (1992) and confirmed by Burrell and Rousseau (1992), adjusted counting leads to a breakdown in the estimation of Lotka's law.

Therefore, in this paper we follow Nicholls's (1989) recommendation of using what is known as the *complete count*, namely, a multiplicative strategy in which any article co-authored by two or more scholars is wholly assigned as many times as necessary to each of them. Of course, this means that the set of articles actually studied increases quite dramatically: the total number of articles in what we call the *double extended count* becomes 48,200,834, or 4.6 times larger than the number of distinct articles, and 2.8 times larger than the 17.2 million authors in the extended count – a fraction approximately equal to the average number of authors per article in the dataset.

Nevertheless, in RCC we study the consequences of adopting an adjusted count for the assignment of individual responsibility to the authors of articles written by two or more scientists. To save space, we will only summarize the main results in Section 5 devoted to extensions.[7]

Next, by way of comparison, we briefly review the characteristics of the dataset used in Ioannidis, Boyack, and Klavans (2014), as well as the way these authors tackle the above methodological problems. To begin with, it should be noted that Ioannidis et al. (2014) use the Scopus database that includes all genres of published items in 1996–2011 among which, nevertheless, journal articles predominate. Instead, our dataset consists only of research articles published in academic journals, excluding publications in local journals, as well as magazine and trade journals.

With regard to the four methodological issues, Ioannidis et al.'s (2014) approach can be summarized as follows. (i) These authors use a classification system – previously developed in Börner et al. (2012) and Boyack and Klavans (2014) – that allocates each paper to a separate scientific discipline. They distinguish between 13 broad fields. (ii) Rather than attempting to disambiguate authors on their own, as we have done, Ioannidis et al. (2014) use Scopus author identifiers. (iii) This contribution approaches the problem of allocating authors to fields in a different way to ours. In the first place, because in Ioannidis et al. (2014) every publication belongs to a single field, the percentage of authors with publications in several fields is expected to be smaller than in our extended count. In any case, each author is allocated to a specific field depending on what is the most common field of the papers he/she has authored. This implies the dismissal of available information (e.g. the smaller contribution of authors to fields that are not their main field). (iv) Finally, nothing is said about the assignment of individual responsibilities in the case of a publication with multiple authors.

The between-field variation of any characteristic will be measured by means of the coefficient of variation (CV hereafter) of the characteristic in question over the 30 fields. The CV is defined as the ratio of the standard deviation over the mean. There is no generally agreed criterion in Statistics concerning when a CV is "large" or "small", possibly because this distinction is context dependent. Although any reader is free to apply a different criterion, in this paper we will use the following convention. We say that the between-field variability of any characteristic is

- "Small", if $CV \leq 0.10$, meaning that the standard deviation of this characteristic over the 30 fields is smaller than or equal to 10% of the mean.
- "Intermediate", if $0.10 < CV \leq 0.30$.
- "Large", if $0.30 < CV \leq 0.60$.
- "Very large", if $CV > 0.60$.

3. Productivity as the number of articles per person

3.1. The size and the mean of productivity distributions

In this section, we define individual productivity as the number of distinct articles written by each individual independently of the number of authors involved. We begin by analyzing two basic characteristics: the size and the mean of field productivity distributions. The information is presented in Table 1.

The following three points should be noted.

1. As expected, the between-field variability of field sizes is very large: the CV over the 30 fields is 1.3, with sizes varying from 43,614 authors in Information & Communication Sciences to 3,258,493 authors in Clinical Medicine (column 1 in Table 1).
2. Taking into account that we study the publication performance of individuals over a period of nine years, field mean productivity values are generally low (column 3 in Table 1). On one hand, researchers in Astronomy & Astrophysics or Physics & Materials Science who, on average, publish 8.2 and 4.3 papers, are seen to publish one article every 9/8.2 = 1.1 or

[7] The interested reader can find in the SMS the following statistical information concerning topics discussed in Section 2: (i) The number of distinct articles assigned by Thomson Reuters to more than one field – an information that clarifies the construction of the extended count (Table I). (ii) Percentage of researchers who have their *oeuvre* in one field in the two cases of authors of articles assigned by Thomson Reuters to a single field, and authors in the extended count (Table II). (iii) The mean and the maximum number of authors per article per field (Table III).

Table 1
Number of authors, mean number of articles per author, and percentage of authors with one publication by scientific field for the entire population in the extended count.

	Number of authors	%	Mean μ_1	% authors with one publication
	(1)	(2)	(3)	(4)
Agriculture & Food Science	495,525	2.9	2.3	68.2
Astronomy & Astrophysics	128,908	0.7	8.2	51.6
Basic Life Sciences	1,889,540	11.0	2.6	63.4
Basic Medical Sciences	406,529	2.4	2.0	69.6
Biological Sciences	785,341	4.6	2.4	66.7
Biomedical Sciences	1,925,259	11.2	2.6	64.8
Chemistry & Chemical Eng.	1,662,043	9.7	3.0	65.5
Civil Eng. & Construction	125,858	0.7	1.8	73.4
Clinical Medicine	3,258,493	18.9	3.1	66.8
Computer Sciences	416,676	2.4	2.2	68.3
Earth Sciences & Technology	388,739	2.3	2.9	63.9
Economics & Business	122,889	0.7	2.3	65.8
Educational Sciences	116,491	0.7	1.6	77.5
Electrical Eng. & Telecomm.	504,441	2.9	2.3	68.3
Energy Sc. & Technology	309,527	1.8	2.3	66.9
Environmental Scs. & Tech.	619,686	3.6	2.3	67.1
General & Industrial Eng.	150,233	0.9	1.7	75.3
Health Sciences	409,315	2.4	2.1	71.5
Information & Comm. Scs.	43,614	0.2	1.6	77.5
Insts. & Instrumentation	226,792	1.3	2.3	68.0
Law and Criminology	53,544	0.3	1.6	77.6
Management & Planning	72,120	0.4	1.7	73.6
Mathematics	205,178	1.2	2.9	60.8
Mechanical Eng. & Aerospace	297,584	1.7	2.1	69.3
Multidisciplinary Journals	376,086	2.2	1.6	73.7
Physics & Materials Science	1,671,513	9.7	4.3	65.1
Psychology	253,346	1.5	2.4	67.7
Social & Behavioral Sciences	74,552	0.4	1.5	77.7
Sociology & Anthropology	90,123	0.5	1.6	75.3
Statistical Sciences	119,488	0.7	2.2	68.4
Total	17,199,433	100.0		
Average	573,314		2.5	69.0
Coefficient of variation	1.3		0.50	0.08

9/4.3 = 2.1 years, respectively. On the other hand, researchers in a number of fields (such as Social & Behavioral Sciences; Information & Communication Sciences; Educational Sciences; Sociology & Anthropology, and perhaps not surprisingly, Multidisciplinary Journals) publish one paper, approximately, only every six years (i.e. 9 years/∼1.5).

However, such low mean values are easily understood when we realize that, on average, about 68% of authors in all fields publish a single article during this nine-year period (column 4 in Table 1). However, as already indicated in note 4, the author name disambiguation algorithm promotes precision over recall. Thus, it should be acknowledge that when there is limited information to cluster the publications of a certain author, the algorithm may occasionally split the *oeuvre* of an author into clusters with only one publication. Future research will focus on the exploration of more refined datasets that do not suffer from this shortcoming.

Finally, for comparison purposes, it should be noted that the percentage of authors publishing a single paper during a 16 year period in Ioannidis et al. (2014) is 58.2%. One possible reason for this percentage to be lower than ours is that, as we saw in Section 2, in this contribution each author is allocated to a specific field depending on what is the most common field of the papers he/she has authored. Therefore, Ioannidis et al. (2014) eliminate all cases of authors with a single publication in an uncommon field which, on the contrary, are all included in our dataset.

3. Thirdly, as expected, the CV of mean productivity (0.50) is rather high, ranging from 1.6 in four fields (Educational Sciences, Information & Communication Sciences, Law & Criminology, and Multidisciplinary Journals) to 4.3 in Physics & Materials Science and 8.2 in Astronomy & Astrophysics.

For our purposes, it is important to keep track of the between-field variability in all dimensions of our analysis. Therefore, for further reference, Table 2 includes the average value and the CV over the 30 fields for some key characteristics of different types of productivity distributions.

In any case, given the between-variability of the size and the mean in field productivity distributions, it is clear that, as indicated in Section 1, we need to pursue the analysis focusing on the distributions' shape using size- and scale-independent techniques – a task to which we devote the next sub-section.

Table 2
Between-field variation. Average and coefficient of variation over the 30 fields of some characteristics of productivity distributions.

	Individual productivity = number of articles per author			
	Size = number of authors	First mean μ_1	Second mean μ_2	Skewness index
	(1)	(2)	(3)	(4)
I. Total population				
Average	573,314	2.5	7.9	∼1
Coefficient of variation	1.3	0.50	0.80	∼0
II. Successful authors with above average productivity				
Average	109,710	7.9	17.0	0.72
Coefficient of variation	1.2	0.80	0.92	0.25
	Individual productivity = mean citation per article per author			
	Size = number of authors	First mean μ_1	Second mean μ_2	Skewness index
	(1)	(2)	(3)	(4)
III. Total population				
Average	573,314	8.6	23.8	0.64
Coefficient of variation	1.3	0.95	1.14	0.10
IV. Successful authors with above average productivity				
Average	162,268	23.8	49.4	0.63
Coefficient of variation	1.3	1.14	1.25	0.08

3.2. The skewness of productivity distributions

Consider a population of N individuals, indexed by $i = 1, \ldots, N$. Assume that we have information about a certain individual characteristic, say x_i for each i; in other words, assume that we have information about the ordered distribution $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_N)$ with $x_1 \leq x_2 \cdots \leq x_i \cdots \leq x_N$. Let $X = \sum_i x_i$ be the total amount of this characteristic in the population. In our application of the CSS technique, two *characteristic scores* are determined: μ_1 = mean of $\boldsymbol{x}$, and μ_2 = mean over the individuals with $x_i > \mu_1$. Next, using the scores we partition the population into three categories: (i) individuals with $x_i \leq \mu_1$; (iii) individuals with $\mu_1 < x_i \leq \mu_2$, and (iii) individuals with $x_i > \mu_2$. The CSS technique allows us to describe distribution $\boldsymbol{x}$ by means of two sets of statistics: the percentages of individuals in the three categories, and the percentages of X attributed to the individuals in each category.

When x_i is the number of publications per author in a certain field, μ_1 is the mean number of publications for the entire productivity distribution, μ_2 is the mean number of publications for authors with a number of articles above μ_1, and X is the total number of publications in the field. We partition the productivity distribution into three categories: (i) authors with low productivity that publish a number of articles smaller than or equal to μ_1; (iii) fairly productive authors, with productivity greater than μ_1 and smaller than or equal to μ_2, and (iii) authors with remarkable or outstanding productivity above μ_2. The average (the standard deviation), and the CV of the six values over the 30 fields appear in row I in Table 3.

The results are remarkable. The research productivity of scientists in every field is determined by a complex set of factors whose study is beyond the scope of this paper. However, the relatively small standard deviations and CVs in row I in Table 3 indicate that field productivity distributions tend to share some fundamental characteristics. Fig. 2 (where fields are ordered according to the percentage of people in category 1) illustrates the similarity of the partition of authors into the three categories in the different fields. Specifically, we find that, on average, 79.3% of all individuals have productivity below μ_1 and account for approximately 40% of all publications, while individuals with a remarkable or outstanding productivity represent 5.9% of the total and account for 35% of all publications. Thus, we can conclude that field productivity distributions are both similar and very highly skewed indeed in the sense that a large proportion of researchers have below average productivity, while a small percentage of them account for a disproportionate amount of all publications.

Compare the CSS approach with the procedure followed by Ioannidis et al. (2014). The latter identifies how many authors have published at least once in each and every year in the 16 year period 1996–2011. These authors are said to have an uninterrupted, continuous presence (UCP) in the scientific literature over this period. There are 150,608 UCP authors in a dataset of 15,153,100 scholars, or 0.99% of the total. Ioannidis et al. (2014) find that, contrary to our results, the relative proportion of UCP authors across scientific disciplines is different than the respective distribution for non-UCP authors.

This difference in results can be explained by two factors. Firstly, recall that we each solve the four methodological problems discussed in this paper in a different way (see Section 2). Secondly, and more importantly, Ioannidis et al. (2014) define the UCP condition equally for all fields. However, fields with a large average number of publications per author will tend to have a larger percentage of UCP authors. In our case, the procedure to partition authors into three categories abstracts from these well-known differences in average productivity across fields. Thus, it should come as no surprise that "*The presence of the UCP pattern is relatively enriched in Medical Research, but also in Mathematics/Physics and Chemistry, while the presence*

Table 3
The skewness of two types of productivity distributions according to the CSS approach. Average, standard deviation, and coefficient of variation over 30 fields of the percentages of individuals, and the percentages of articles (or citations) by category.

	Individual productivity = number of articles per author					
	Percentage of people in category			Percentage of articles in category		
	1	2	3	1	2	3
I. Total population						
Average (Std. dev.)	79.3 (3.4)	14.8 (2.4)	5.9 (1.2)	40.4 (7.0)	24.5 (1.8)	35.1 (6.3)
Coeff. of variation	0.04	0.17	0.19	0.17	0.07	0.18
II. Successful authors with above average productivity						
Average (Std. dev.)	71.4 (2.4)	19.8 (1.7)	8.8 (1.1)	41.4 (7.0)	27.4 (1.5)	31.1 (3.5)
Coeff. of variation	0.03	0.09	0.12	0.10	0.06	0.11
	Individual productivity = mean citation per article per author					
	Percentage of people in category			Percentage of total citations in category		
	1	2	3	1	2	3
III. Total population						
Average (Std. dev.)	71.0 (2.1)	20.7 (1.2)	8.3 (1.1)	22.6 (3.1)	40.2 (3.7)	37.2 (4.6)
Coeff. of variation	0.03	0.06	0.13	0.14	0.09	0.12
IV. Successful authors with above average productivity						
Average (Std. dev.)	71.0 (2.2)	20.3 (1.0)	8.3 (1.2)	52.0 (5.0)	27.7 (1.8)	20.3 (3.7)
Coeff. of variation	0.03	0.06	0.13	0.10	0.06	0.18

Total population, row I (same interpretation for row III substituting m_1 and m_2 for μ_1 and μ_2).
Category 1 = people with a low productivity, below μ_1 (mean productivity).
Category 2 = people with a fair productivity, above μ_1 and below μ_2 (mean productivity of people with productivity above μ_1).
Category 3 = people with a remarkable or outstanding productivity, above μ_2.
Successful population, row II (same interpretation for row IV substituting m_3 for μ_3).
Category 1 = people with a fair productivity, between μ_1 and μ_2.
Category 2 = people with a remarkable productivity, between μ_2 and μ_3.
Category 3 = people with outstanding productivity, above μ_3.

of the non-UCP pattern is relatively enriched in Social Sciences and Humanities (the UCP pattern is practically non-existent in the Humanities), as well as Engineering and Computer Sciences/Electrical Engineering." (p. 4).

On the other hand, as indicated in Section 1 we have also considered the computation of numerical skewness indexes for all fields. The problem, of course, is that extreme observations with a very large number of citations are known to be prevalent in citation distributions (see inter alia Herranz & Ruiz-Castillo, 2012a; Li & Ruiz-Castillo, 2014). This presents a challenge for conventional measures of skewness that are very sensitive to extreme observations. Fortunately, robust measures of skewness based on quartiles have been developed in the statistics literature. In particular, given a process $\{y_t\}$, $t = 1, \ldots, T$, where the y_t's are independent and identically distributed with a cumulative distribution function F, the Groeneveld and Meeden (1984) robust measure is

$$GM = \frac{\mu - Q_2}{E|y_t - Q_2|},$$

where $Q_2 = F^{-1}(0.5)$ is the second quartile of y_t, or the median of the distribution, and the expectation in the index denominator is estimated by the sample mean of the deviations from the median in absolute value.

For the interpretation of results, the following three properties should be taken into account. Firstly, like the CSS approach, the *GM* index is scale- and size-independent. Secondly, whenever the mean is greater than the median – as it is always the case in our dataset – the *GM* index takes positive values. Thirdly, the *GM* index is bounded in the interval $[-1, 1]$. However, whenever the process consists of natural numbers and the lower 50% of the observations are equal to, say, a value z, then the median is z, and the *GM* index reaches its upper bound of 1. Thus, extreme distributions of this sort drives the *GM* index to its upper bound. As we will see presently, this is a useful property to have in our case.

Recall that, the percentage of authors with a single publication in the period 2003–2011 is greater than 50% in all fields (column 4 in Table 1). Therefore, the median and the *GM* index are equal to 1 in all cases. This clearly reinforces the idea that we are facing an unusual situation in which field productivity distributions are extremely skewed, as well as very similar to each other.

3.3. Successful authors

So far we have studied the productivity of all authors measured by the number of their publications in the 2013–2011 period. Given the high percentage of authors with a single publication, the low values of μ_1, and the fact that for the 30

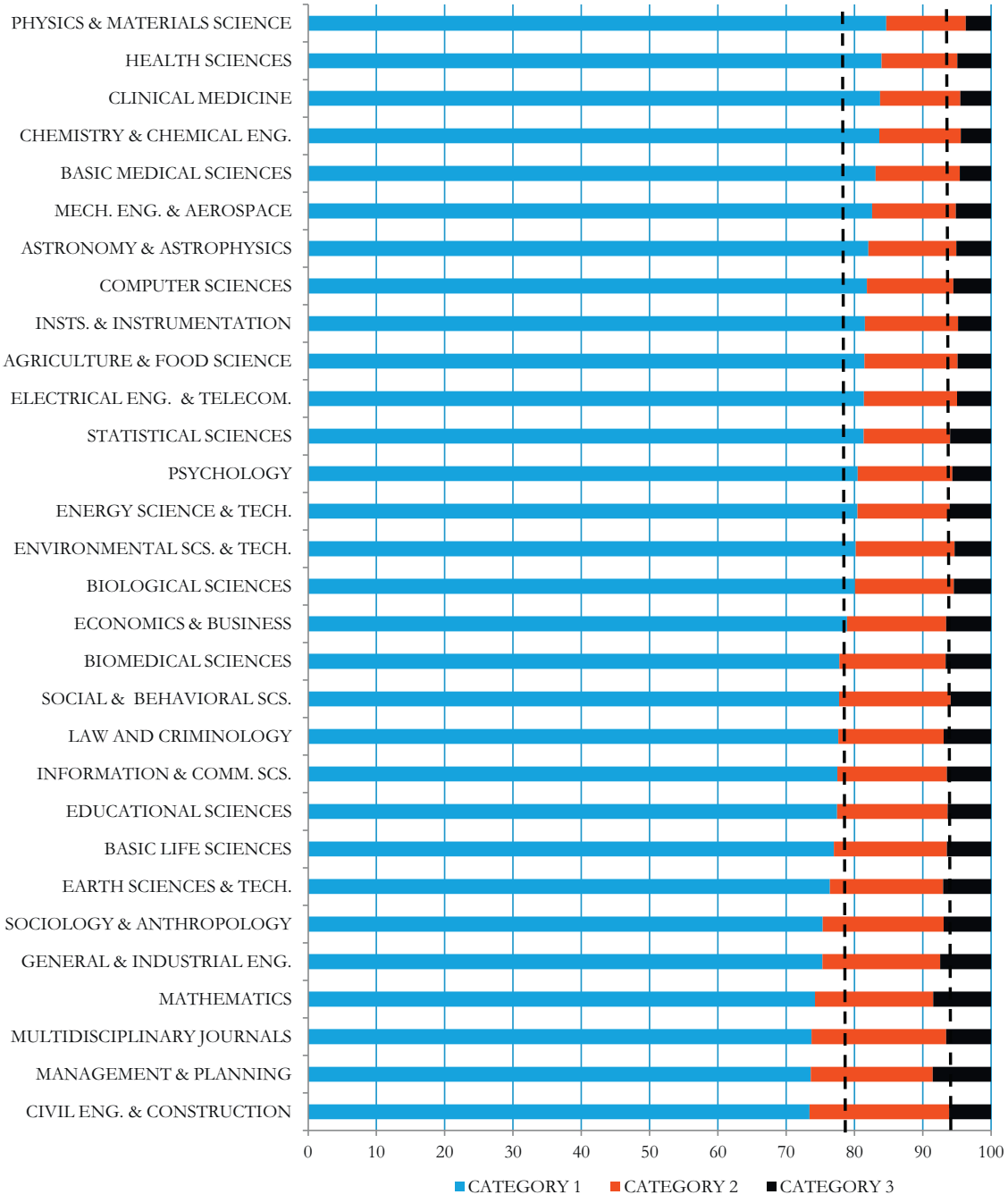

Fig. 2. Partition of productivity distributions into three categories according to the CSS technique. Productivity = number of articles per person. Population as a whole.

fields the *GM* index reaches its upper bound (i.e. 1), it seems interesting to study the behavior of relatively productive authors. Specifically, we define *successful authors* as those with above average productivity in their respective field. Their total number is 3,291,299, or 19.1% of the population as a whole. The following two points should be emphasized. Firstly, the distribution of successful authors by field, which appears in columns 1 and 2 in Table A in the Appendix, is not very different from the distribution for the population as a whole. Secondly, the between-field variability of field sizes is again very high: the coefficient of variation is 1.2 (row II and column 1 in Table 2).

In order to apply the CSS approach, we need two means: μ_2, which has been already introduced, and the mean productivity μ_3 with productivity above μ_2. The information concerning μ_2 and μ_3 for all fields is in columns 3 and 4 in Table A in the

Appendix. As expected, the coefficients of variation over all fields for μ_2 and μ_3 are very high (row II, columns 2 and 3 in Table 3).

As far as the *GM* index is concerned (column 5 in Table A in the Appendix), median productivity for successful authors is equal to two in six fields. Consequently, their *GM* index reaches again its upper bound. The lowest values are 0.41 and 0.47 for Management & Planning and Economics & Business. For the remaining 12 fields, *GM* ranges from 0.52 to 0.97. Thus, we conclude that most field productivity distributions are again highly skewed according to the *GM* criterion. However, the coefficient of variation of *GM* values is 0.25 (row II, column 5 in Table 2) – a relatively small value in comparison with the coefficients of variation for field size and mean productivity in columns 1–3 and row II in Table 2.

A summary of results from the CSS approach for successful authors is presented in row II in Table 3. The comparison with the population as a whole (row I in Table 3) yields three very interesting results. Firstly, all standard deviations and coefficients of variation are smaller in row II than in row I, indicating that field productivity distributions are now more similar than before. The situation for the percentage of successful authors in the three categories is illustrated in Fig. 3 (where fields are ordered according to the percentage of people in category 1). Secondly, on average, the percentage of people in category 1 (with productivity below μ_2) is eight points smaller than before (with productivity below μ_1). Furthermore, the percentage of successful people in categories 2 and 3 is five and three points greater than for the population as a whole. This agrees with the results obtained with the *GM* criterion: field productivity distributions are still highly skewed, but the degree of skewness is considerably lower than the extraordinary high levels reached for the entire population in each field. Thirdly, relative to the previous situation, the percentage of publications accounted for by the three categories remains essentially constant. Thus, on average, 71.4% of all successful individuals in category 1 account for approximately 41.4% of all publications, while researchers in category 3 represent 8.8% of the total and account for 31.1% of all publications.[8]

4. Productivity as the mean citation per article per person

4.1. Characteristics of productivity distributions

As indicated in Section 1, measuring productivity as the number of publications per author in a certain period has a long history in Scientometrics. However, in the dataset used in this paper it is possible to take into account each author's citation impact. Therefore, in this section, we define individual productivity as the mean citation per article per person during 2003–2011. The two measures are essentially uncorrelated (the correlation coefficient between them in the entire sample is 0.02). This reveals that the most prolific authors need not necessarily be those with the highest impact. Thus, the two productivity concepts are best treated separately.

The information concerning the mean and the *GM* index, which can be found in the first two columns in Table B in the Appendix, can be summarized in the following two points. Firstly, mean productivity varies widely, ranging from 3.2 and 3.7 in Mathematics and Computer Sciences, to 13.3 or 15 in Clinical Medicine and Basic Life Sciences. Not surprisingly, the highest value, 49.3, is reached in Multidisciplinary Journals (column 1 in Table B). As usual for mean productivity variables, the coefficient of variation is very high: 0.95 (row III and column 2 in Table 2). Secondly, according to the *GM* index, productivity distributions are highly skewed in all fields: this measure ranges from 0.54 in Basic Medical Sciences to 0.80 in Multidisciplinary Journals. However, for our purposes, it is important to emphasize that the between-field variability of the *GM* index is very low (row III and column 5 in Table 2), indicating considerable similarity across all fields around an average *GM* value of 0.64.

When the variable x_i, introduced in the description of the CSS approach in Section 3.2, is the mean citation per article per author in a certain field, m_2 is the mean productivity for authors with mean citation per article above m_1. Consider again the partition of the distribution into three broad classes: (i) authors with productivity smaller than or equal to m_1; (ii) authors with productivity between m_1 and m_2, and (iii) authors with a remarkable or outstanding productivity above m_2. In this case, the variable X is the sum of the values of the variable "mean citation per article" over all authors in the field. However, we find it more interesting to inform about the allocation of the *total number of citations* into the three categories. The average (the standard deviation), and the CV over the 30 fields of the percentage of authors in the three categories, as well as the corresponding percentages of the total number of citations accounted for by each one appear in row III in Table 3. The main result is that the CVs in row III are generally very low. Fig. 4 (where fields are ordered according to the percentage of people in category 1) illustrates the similarity of the partition of authors into the three categories in all fields. Moreover, the comparison of results in rows III and IV in Table 3, as well as Figs. 3 and 4, indicate that the between-field variability of skewness in the CSS approach under the second productivity definition is comparable to the between-field variability for successful authors under the first definition.

[8] The SMS includes information on the following issues: (i) The comparison of the distribution of authors in the extended count with the distribution of articles in the extended count and the double extended count (Table IV). (ii) CSS results at the field level for the entire population and successful authors when individual productivity is defined as the number of articles per author (Tables V and VI).

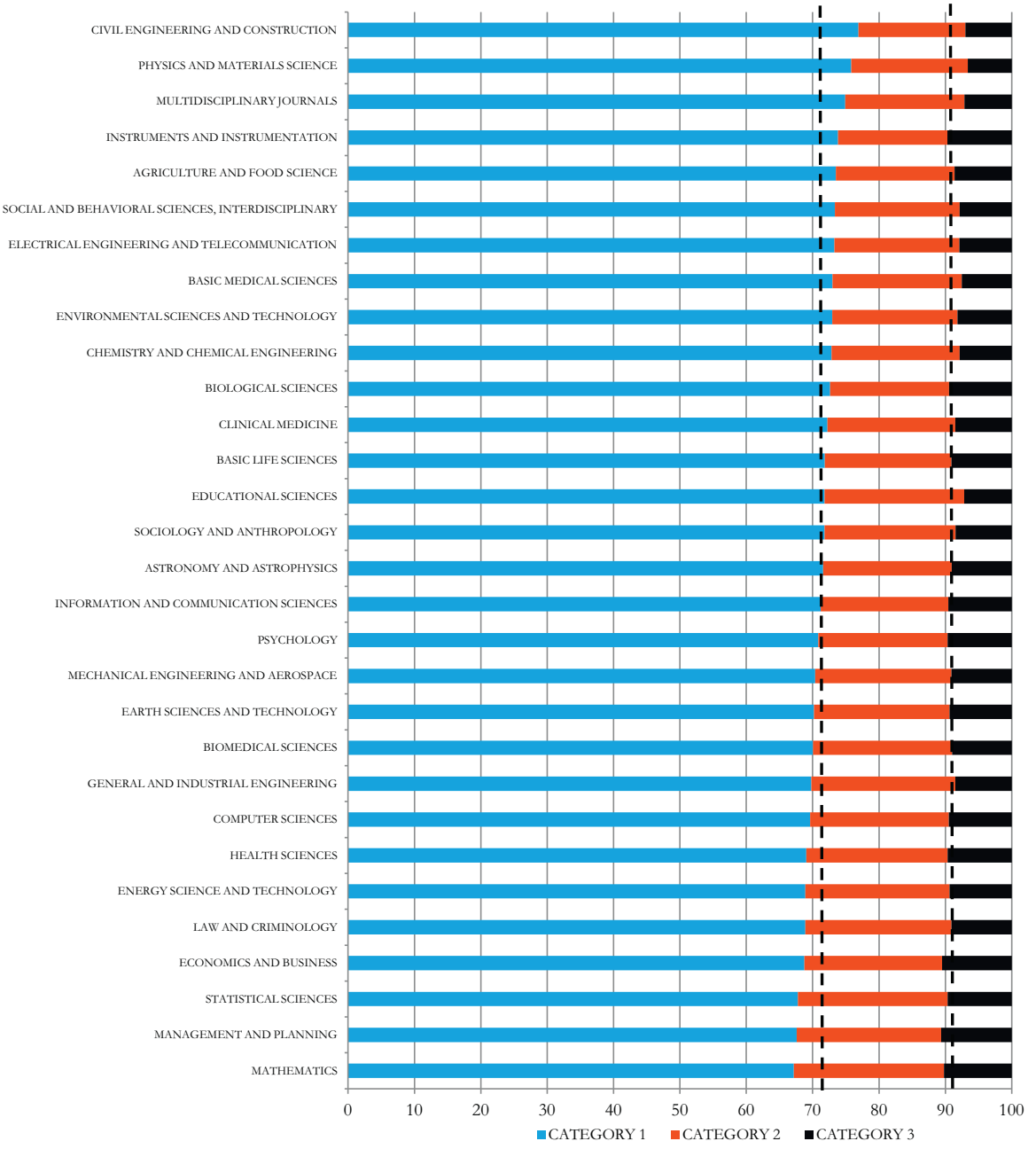

Fig. 3. Partition of productivity distributions into three categories according to the CSS technique. Productivity = number of articles per person. Successful authors with above average productivity.

4.2. Successful authors

Just as before, it is interesting to study successful authors, namely, scientists with above average productivity. Their total number is 4,868,030, or 28.3% of the population as a whole.[9] The key characteristics of the field productivity distributions for successful authors, which are presented in Table B in the Appendix, can be summarized in the following three points. Firstly, the field size distribution is not very different from the distribution for the population as a whole. As a matter of fact, the between-field variability of field size is exactly as high as before. Secondly, as expected, mean citations per article are very

[9] The two productivity measures are again essentially uncorrelated (the correlation coefficient between the number of publications and the mean citation per article for the 4,868,030 successful authors is – 0.03).

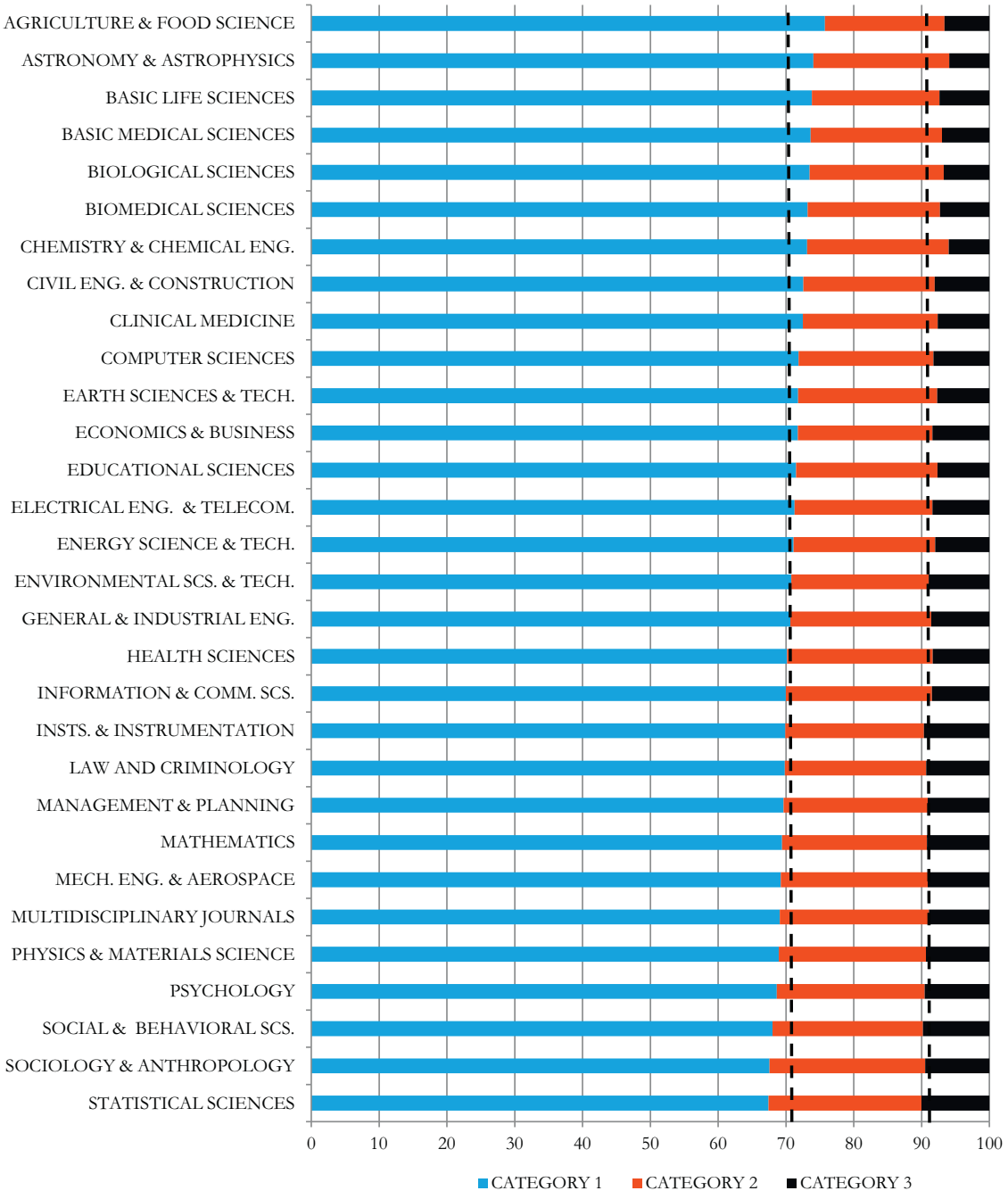

Fig. 4. Partition of productivity distributions into three categories according to the CSS technique. Productivity = mean citation per article per person. Population as a whole.

different across fields. Thirdly, on the contrary, the between-field variability of the skewness of productivity distributions according to the *GM* index is extremely low: the CV is 0.08 (columns 1–4 in row IV in Table 2).

A summary of results from the CSS approach is presented in row IV in Table 3. The comparison with the population as a whole (row III in Table 3) yields a first fundamental result: on average, the partition of both populations over the three CSS categories is essentially the same. Furthermore, judging from the CVs, the similarity across fields is also the same as before. This clearly illustrates the fractal nature of the skewness of productivity distributions when productivity is measured as the mean citation per article per person (for a graphical illustration, compare Figs. 4 and 5). On the other hand, it is worth emphasizing that the results for successful authors according to both productivity definitions are practically the same (compare rows II and IV in Table 3, as well as Figs. 3 and 5).

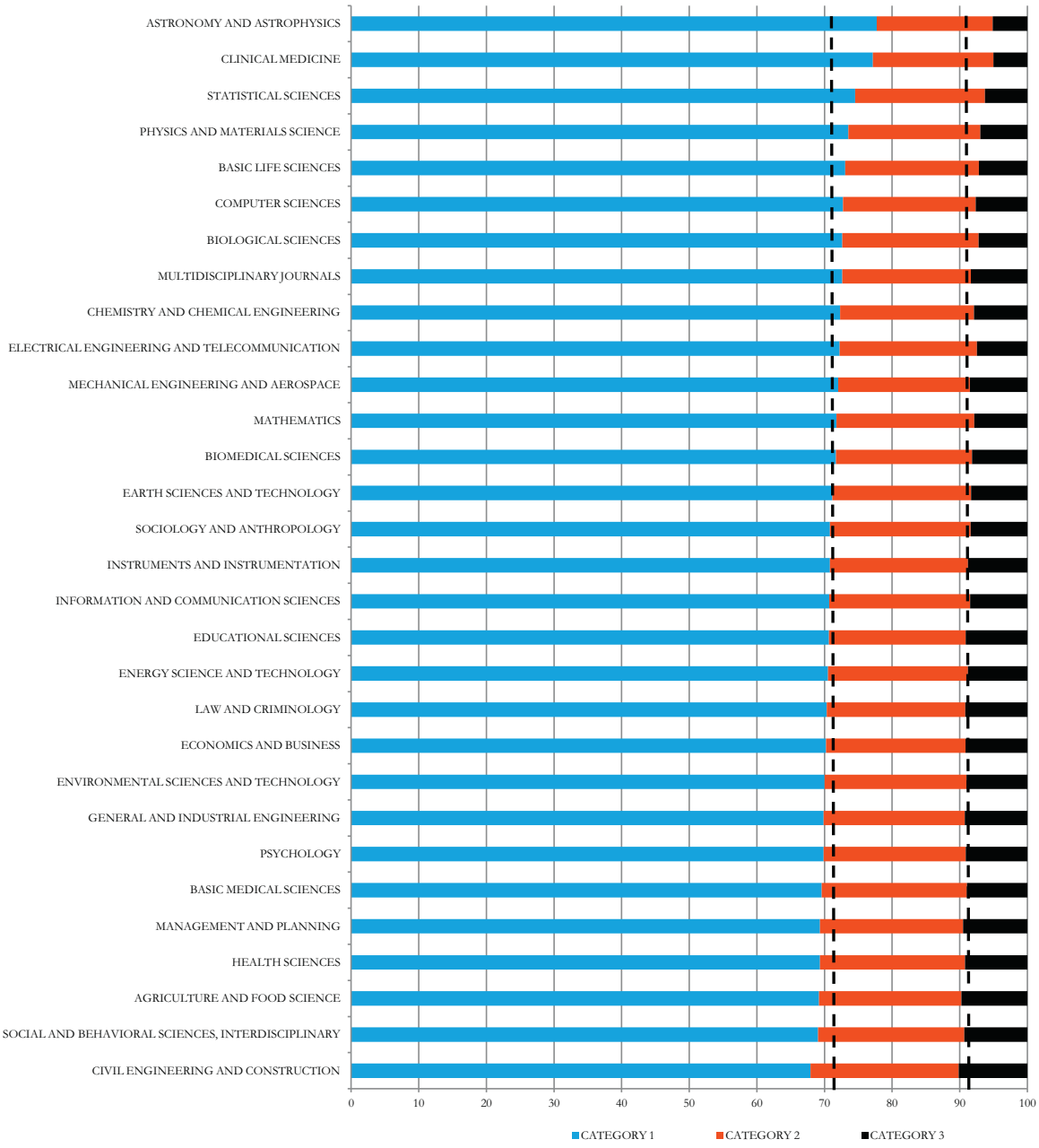

Fig. 5. Partition of productivity distributions into three categories according to the CSS technique. Productivity = mean citation per article per person. Successful authors with above average productivity.

We close this section by noting the following two minor points. Firstly, in the right-hand side of row III in Table 3, category 2 accounts for a large percentage of total citations (40.2%). The explanation lies in the fact that this category includes authors with a relatively large number of publications per capita – on average for all fields, the mean number of articles per person in categories 1, 2, and 3 is 2.1, 3.5, and 2.6, respectively – combined with the high correlation between the number of publications and the total citations received. Secondly, since category 2 in the population as a whole becomes category 1 in the subset of successful authors, the latter accounts again for a large percentage of total citations (52.0%) in row IV. In turn, the explanation of the sequence of percentages of total citations in this row is that less productive among successful authors according to the second productivity definition have relatively many publications, and hence, plenty of total citations, while truly productive authors in terms of mean citations per article have fewer but highly cited publications that, nevertheless, account for a relatively low percentage of total citations. This is, of course, consistent with the very small, but negative

coefficient of correlation between the number of publications and the mean citation per article among successful authors documented in note 12.[10]

5. Some extensions of this study

In this section we summarize some of the results presented in RCC that are relevant for this study.

1. We have systematically studied a second size- and scale-independent characteristic of productivity distributions, namely, field productivity inequality. In an analogous way to the CSS approach in the skewness case, we could have compared field productivity inequality using Lorenz curves, a statistical construction extensively used in Economics for income distributions that has also been used in Scientometrics (see inter alia Rousseau, 2011, and other references cited there). Similarly, given that extreme observations may strongly affect any productivity inequality scalar measure, in an analogous way to the *GM* skewness index we could have used a robust inequality index. However, for brevity, in RCC we have only used the coefficient of variation as our productivity inequality index.

   Intuitively, large individual variability in authors' productivity within each field should yield high values of any productivity inequality index we care to use. Indeed, this is the first result that we obtain in RCC. (The average of the coefficient of variation over the 30 fields ranges from 1.05 to 2.14 for the entire population and for successful authors according to both productivity definitions). The second result is that the between-field variability of productivity inequality according to the CV is considerably smaller than the between-field variability of field size and field mean productivity. (The CV of the coefficients of variation over the 30 fields for the entire population and successful authors when productivity is measured as the number of publications per author is 0.31 and 0.18, and for the second productivity definition the CV is 0.29 and 0.18.) Coefficients of variation at the field level for the entire population according to the two productivity definitions can be found in SMS Table IX.
2. It could be argued that we should investigate the robustness of our results to an adjusted or fractional approach to the treatment of articles co-authored by two or more persons. As far as the CSS approach is concerned, there are four cases corresponding to the population as a whole and successful authors for each of the two productivity definitions. The conclusion is inescapable: in the four cases, the skewness of productivity distributions in each field, and the similarity of productivity distributions across fields when using the complete or the adjusted approach are essentially the same (Tables X–XIII contains the numerical results obtained in Section 5 in RCC).
3. Finally, RCC includes a comparison of our results for the total population (rows I and III in Table 3 in RCC) with those obtained in previous research in two cases: the results obtained in Kyvic (1989), and those concerning citation distributions, namely, the distributions of the number of citations received by articles published in a certain period (Albarrán, Crespo, Ortuño, & Ruiz-Castillo, 2011; Albarrán & Ruiz-Castillo, 2011; Li, Radicchi, Castellano, & Ruiz-Castillo, 2013; Radicchi & Castellano, 2012; Radicchi, Fortunato, & Castellano, 2008, Waltman, Van Eck, & Van Raan, 2012).[11]

6. Summary and further research

This paper has exploited a unique large dataset consisting of 7.7 million distinct articles published in the period 2003–2011 in academic journals, with a variable citation window from the publication year until 2012. We had to overcome the obstacles posed by four methodological problems: the multiple assignment by Thomson Reuters of articles to multiple journal subject categories, for which we followed a multiplicative approach; the identification of authors in articles, which we solved applying a novel author disambiguation algorithm; the allocation of authors to fields, for which we follow the rule that researchers who write articles in several fields should be treated as independent, different authors in their respective fields[12]; and the definition of the individual contribution to an article in the case of multiple authorship, for which we also followed a multiplicative approach. After coping with these problems, we end up with a final dataset consisting of 17.2 million authors classified into 30 broad fields.

We have measured individual productivity in two ways that are essentially uncorrelated in our dataset: the number of articles per person, and the mean citation per article per person. In both cases, we have studied, not only the entire

[10] The SMS includes information on the following issues: CSS results at the field level for the entire population and successful authors when individual productivity is defined as the mean citation per article per person (Tables VII and VIII).

[11] The SMS includes information on the following issues: (i) Coefficients of variation at the field level for the entire population according to the two productivity definitions (Table IX). (ii) CSS results at the field level in the fractional approach for the entire population and successful authors when individual productivity is defined as the mean citation per article per person (Tables X and XI), and when individual productivity is defined as the mean citation per article per person (Tables XII and XIII).

[12] The problem of the assignment of authors to fields is a relatively under researched problem, at least from a quantitative point of view. Even when we consider authors of articles assigned by Thomson Reuters to a single field, we find that one third of researchers have their oeuvre in several fields, introducing the idea that publication in different fields (even broad fields as considered in this study) is quite common among scholars. Given the assignment of articles to multiple fields that characterizes Thomson Reuters datasets, the percentage of scholars with activity in more than one field increases to, approximately, two thirds of the 17.2 million authors in the final dataset, a magnitude that is exaggerating the true extent of the phenomenon.

population, but also what we call successful authors, defined as researchers with a number of articles above the mean in their own field.

The main result of the paper is that the skewness and productivity inequality of field productivity distributions is very similar across fields for all samples. In particular, except for the entire population when productivity is measured as the number of publications per person, in the remaining three samples the percentage of scholars that have a low, fair, or outstanding productivity is of the same order of magnitude. It should be added that all these results are robust to the treatment of articles co-authored by two or more persons following a fractional approach.

The results thus summarized are useful to devise the following research strategy for the future. Firstly, rather than a set of models for different types of sciences, we need a single explanation of within-field variation of scientists' productivity. Secondly, the between-field mean productivity differences in our dataset can be attributed to idiosyncratic differences in production and citation practices. However, just as the similarities between field citation distributions at different aggregation levels have recently paved the way for meaningful comparisons of citations for articles belonging to heterogeneous fields (Crespo et al., 2014; Crespo, Li, & Ruiz-Castillo, 2013; Li et al., 2013), the similarities documented in this paper between field productivity distributions open the possibility of establishing meaningful comparisons of productivity for authors belonging to heterogeneous fields. Thirdly, in this paper we have studied the size and the mean of individual citation distributions. For authors with a minimum number of articles in each field, we could investigate other size- and scale-independent characteristics at the individual level, such as citation inequality and citation skewness. This analysis leads to investigating the possibility of accounting for the characteristics of citation distributions at the field level in terms of the characteristics of citation distributions at the individual level.

In addition, within the methodological framework defined in Section 2 there are two issues for further research. Firstly, it would be relevant to investigate whether the productivity distributions studied in this paper follow a simple functional form. In the case of the number of publications per person, this exercise should start by verifying whether productivity distributions satisfy the generalized Lotka's law. Secondly, it would be interesting to study the distribution of individual mean citations *conditional* on the number of publications in each field. Since, as we have seen, the number of publications and the mean citation per article per person are largely uncorrelated in every field, the conjecture is that conditional distributions are very similar to the marginal distribution, that is, to the distribution of the mean citation per article per person studied in Section 4.

Finally, among the extensions that involve methodological changes or new pieces of information, we mention the following four.

1. It would be important to study the robustness of our results using a classification system where every article is assigned to a single scientific field – a possibility is the publication-level algorithmic methodology introduced by Waltman and Van Eck (2012), and further studied in Ruiz-Castillo and Waltman (2014). In view of the discussion in Section 2, this would tend to reduce the degree in which scholars appear as authors in several fields. On the other hand, the new classification system will typically consist of a similar number of broad fields that, however, would be quite different to the ones we have studied here. Thus, we could test the robustness of our results to a change in the set of fields considered.
2. We should study the issues researched in this paper using an author name disambiguation algorithm different from Caron and van Eck (2014).
3. So far, we have studied a rich dataset informing about publications, authors, and citations during a nine year period. However, as in most of the studies in the productivity literature, we do not have information concerning authors' ages. This poses two problems. Firstly, because of age and/or cohort effects our measures of productivity for authors of different ages and/or cohorts are not actually comparable. Secondly, some young (old) people are only observed during a reduced number of years at the end (beginning) of the period. Consequently, even in the absence of age and cohort effects, the censored productivity measures of these people are not comparable with the productivity of scientists keeping on publishing during the entire period.[13]
4. In this paper we have studied a large set of authors in a number of fields who publish their research during a fixed, relatively short period of nine years. It would be very interesting to follow the publication dynamics of authors in different fields over their entire research career.

Acknowledgements

This paper was conceived while Ruiz-Castillo enjoyed the hospitality of the Centre for Science and Technology Studies, Leiden University, The Netherlands, during the 2013 spring term. Ruiz-Castillo also acknowledges financial help from the Spanish MEC through grant ECO2010-19596. The authors acknowledge fruitful conversations with Raquel Carrasco, comments from the participants in a CWTS seminar, and two referee reports. All remaining shortcomings are the sole responsibility of the authors.

[13] As emphasized by Wagner-Döbler (1995), and Wagner-Döbler and Berg (1995), rather than a study of the varying intensity with which scientists contribute in their respective fields, what we have accomplished with our cross-section of authors of different ages is a bibliometric description, a "snapshot" of the state of the different fields with regard to the structure of scientific participation.

Appendix.

Table A
Number of authors, second and third mean productivities, and *GM* skewness index for successful authors when individual productivity is defined as number of articles per author.

	Number of authors (1)	% (2)	Second mean μ_2 (3)	Third mean μ_3 (4)	Skewness index (5)
Agriculture & Food Science	91,920	2.8	7.3	15.8	0.58
Astronomy & Astrophysics	23,158	0.7	36.1	82.7	0.66
Basic Life Sciences	435,390	13.2	7.5	15.9	0.63
Basic Medical Sciences	68,827	2.1	6.4	13.0	0.77
Biological Sciences	156,580	4.8	7.1	14.9	0.58
Biomedical sciences	427,560	13.0	7.9	16.3	0.66
Chemistry & Chemical Eng.	271,829	8.3	11.8	27.0	0.69
Civil Eng. & Construction	33,448	1.0	4.1	9.4	0.54
Clinical Medicine	529,675	16.1	12.6	28.6	0.74
Computer Sciences	75,717	2.3	6.8	13.3	0.81
Earth Sciences & Technology	91,723	2.8	8.6	18.1	0.73
Economics & Business	25,911	0.8	6.4	11.5	0.47
Educational Sciences	26,261	0.8	3.6	7.0	1.00
Electrical Eng. & Telecomm.	93,917	2.9	7.4	16.2	0.60
Energy Sc. & Technology	60,589	1.8	6.8	13.3	0.53
Environmental Scs. & Tech.	122,762	3.7	7.2	15.1	0.58
General & Industrial Eng.	37,099	1.1	3.8	7.3	1.00
Health sciences	65,626	2.0	6.9	13.6	0.54
Information & Comm. Scs.	9815	0.3	3.6	6.9	1.00
Insts. & Instrumentation	41,899	1.3	7.4	16.6	0.60
Law and Criminology	11,984	0.4	3.9	7.5	1.00
Management & Planning	19,047	0.6	3.7	6.5	0.41
Mathematics	52,892	1.6	8.0	15.5	0.69
Mechanical Eng. & Aerospace	51,799	1.6	6.7	13.3	0.80
Multidisciplinary Journals	98,931	3.0	3.3	6.3	0.97
Physics & Materials Science	256,345	7.8	20.4	55.2	0.76
Psychology	49,450	1.5	7.6	15.9	0.64
Social & Behavioral Sciences	16,607	0.5	3.3	6.3	1.00
Sociology & Anthropology	22,225	0.7	3.4	6.4	1.00
Statistical Sciences	22,313	0.7	6.9	13.0	0.54
Total	3,291,299	100.0			
Average	109,710.0	3.3	7.9	17.0	0.72
Coefficient of variation	1.2	1.2	0.80	0.92	0.25

Table B
Mean productivity, and *GM* index for the population as a whole, as well as number of authors, second and third mean, and *GM* index for successful authors when individual productivity is defined as mean citation per article per author.

	Total population		Successful authors					
	Mean m_1 (1)	GM index (2)	Number of authors (3)	% authors (4)	Coefficient of variation (5)	Second mean m_2 (6)	Third mean m_3 (7)	GM index (8)
Agriculture & Food Science	7.9	0.6	161,463	3.3	1.00	18.6	33.9	0.59
Astronomy & Astrophysics	10.6	0.65	34,668	0.7	1.19	29.3	72.8	0.72
Basic Life Sciences	15	0.63	538,677	11.1	1.42	39	82.2	0.66
Basic Medical Sciences	8.7	0.54	127,433	2.6	1.08	20.8	38.4	0.61
Biological Sciences	9.2	0.61	221,554	4.6	1.84	24.4	50.2	0.66
Biomedical Sciences	11.8	0.62	578,264	11.9	1.15	29.2	57.3	0.65
Chemistry & Chemical Eng.	9.6	0.59	479,962	9.9	1.23	25.1	50.9	0.65
Civil Eng. & Construction	5.5	0.54	37,900	0.8	1.32	14.4	26.2	0.55
Clinical Medicine	13.3	0.67	846,112	17.4	1.42	38.2	93.8	0.74
Computer Sciences	3.7	0.79	111,769	2.3	1.91	11.3	24.6	0.7
Earth Sciences & Technology	7.4	0.57	120,210	2.5	1.3	18.2	34.7	0.63
Economics & Business	6.3	0.62	34,742	0.7	1.33	17.4	33.6	0.63

Table B (*Continued*)

	Total population		Successful authors					
	Mean m_1 (1)	GM index (2)	Number of authors (3)	% authors (4)	Coefficient of variation (5)	Second mean m_2 (6)	Third mean m_3 (7)	GM index (8)
Educational Sciences	5.2	0.72	32,779	0.7	1.36	14.3	27.9	0.62
Electrical Eng. & Telecom.	4.4	0.62	138,853	2.9	1.57	12.6	26.2	0.7
Energy Science & Technology	6.2	0.6	84,927	1.7	1.42	17.5	34.5	0.63
Environmental Scs. & Tech.	8.2	0.62	188,231	3.9	1.04	20.3	37.5	0.58
General & Industrial Eng.	4.5	0.66	45,926	0.9	1.27	11.4	21.5	0.64
Health Sciences	7.7	0.6	131,002	2.7	1.04	18.4	33.4	0.56
Information & Comm. Sciences	5.7	0.65	12,539	0.3	1.73	15.6	31	0.59
Instr. & Instrumentation	5	0.69	66,619	1.4	1.65	13.6	27.6	0.65
Law and Criminology	4.7	0.67	16,469	0.3	1.16	12.1	22.9	0.55
Management & Planning	6.7	0.65	21,064	0.4	1.38	17.8	33.8	0.57
Mathematics	3.2	0.77	53,750	1.1	1.55	9.7	20.2	0.67
Mechanical Eng. & Aerospace	4.6	0.66	88,833	1.8	1.15	12	23.8	0.61
Multidisciplinary Journals	49.3	0.79	91,198	1.9	1.75	162.7	362	0.68
Physics & Materials Science	7.3	0.68	440,974	9.1	1.46	21.5	47.1	0.66
Psychology	8.6	0.6	78,723	1.6	1.04	21.2	39.4	0.6
Social & Behavioral Sciences	7.3	0.55	22,485	0.5	1.18	18.7	34.4	0.55
Sociology & Anthropology	5.8	0.59	29,235	0.6	1.23	14	26.5	0.62
Statistical Sciences	5.2	0.71	31,669	0.7	1.52	15.1	33.5	0.65
Total			4,868,030	100				
Average	8.6	0.64	162,268	3.3	1.36	23.8	49.4	0.63
Coefficient of variation	0.95	0.1	1.3	1.3	0.18	1.14	1.25	0.08

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.joi.2014.09.006.

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.

Albarrán, P., Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2014). Differences in citation impact across countries. *Journal of the American Society for Information Science and Technology*, http://dx.doi.org/10.1002/asi.23219 (in press)

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.

Alvarado, R. (2012). La colaboración de los autores en la literature producida sobre la Ley de Lotka. *Ciência da Informação*, *40*, 266–279.

Böner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Larivière, V., et al. (2012). Design and update of a classification system: The UCSD map of science. *PLoS ONE*, *7*, e39464.

Boyack, K. W., & Klavans, R. (2014). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the American Society for Information Science and Technology*, http://dx.doi.org/10.1002/asi.22990

Burrell, Q., & Rousseau, R. (1992). Breakdown of the robustness property of Lotka's law. *Journal of the American Society for Information Science*, *46*, 97–102.

Buter, R. K., & van Raan, A. F. J. (2011). Non-alphanumeric characters in titles of scientific publications: An analysis of their occurrence and correlation with citation impact. *Journal of Informetrics*, *5*, 608–617.

Caron, E., & van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In E. Noyons (Ed.), *19th International Conference on Science and Technology Indicators. "Context counts: Pathways to master big data and little data"* (pp. 79–86). Leiden: CWTS-Leiden University.

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

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.

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.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, *52*, 558–569.

De Solla Price, D. (1963). *Little science, big science*. New Haven, CT: Yale University Press.

Groeneveld, R. A., & Meeden, G. (1984). Measuring skewness and kurtosis. *The Statistician*, *33*, 391–399.

Herranz, N., & Ruiz-Castillo, J. (2012a). Multiplicative and fractional strategies when journals are assigned to several sub-fields. *Journal of the American Society for Information Science and Technology*, *63*, 2195–2205.

Herranz, N., & Ruiz-Castillo, J. (2012b). Sub-field normalization procedures in the multiplicative case: Average-based citation indicators. *Journal of Informetrics*, *6*, 543–556.

Herranz, N., & Ruiz-Castillo, J. (2012c). Sub-field normalization procedures in the multiplicative case: High- and low-impact citation indicators. *Research Evaluation*, *21*, 113–125.

Herranz, N., & Ruiz-Castillo, J. (2013). The end of the European paradox. *Scientometrics*, *95*, 453–464.

Hoekman, J., Frenken, K., & Tijssen, R. J. W. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy*, *39*, 662–673.

Hoekman, J., Scherngell, T., Frenken, K., & Tijssen, R. J. W. (2013). Acquisition of European research funds and its effect on international scientific collaboration. *Journal of Economic Geography*, *13*, 23–52.

Ioannidis, J. P. A., Boyack, K., & Klavans, R. (2014). Estimates of the continuously publishing core in the scientific workforce. *PLOS ONE*, *9*, e101698.

Kyvic, S. (1989). Productivity differences, fields of learning, and Lotka's law. *Scientometrics*, *15*, 205–214.
Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology*, *63*, 1030–1047.
Li, Y., Radicchi, F., Castellano, C., & Ruiz-Castillo, J. (2013). Quantitative evaluation of alternative field normalization procedures. *Journal of Informetrics*, *7*, 746–755.
Li, Y., & Ruiz-Castillo, J. (2014). The impact of extreme observations in citation distributions. *Research Evaluation*, *23*, 174–182.
Lindsey, D. (1980). Production and citation measures in the sociology of science: The problem of multiple authorship. *Social Studies of Science*, *10*, 145–162.
Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science*, *16*, 317–323.
Nicholls, P. T. (1989). Bibliometric modeling processes and the empirical validity of Lotka'a law: The case of adjusted counts for multiauthorship attribution. *Journal of the American Society for Information Science*, *43*, 645–647.
Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2014). Within and across department variability in individual productivity. The case of economics. *Scientometrics*, http://dx.doi.org/10.1007/s11192-014-1449-6 (in press)
Radicchi, F., & Castellano, C. (2012). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS ONE*, *7*, e33833.
Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 17268–17272.
Rousseau, R. (1992). Breakdown of the robustness property of Lotka's law. *Journal of the American Society for Information Science*, *31*, 21–39.
Rousseau, R. (2011). Lorenz curves determine partial orders for comparing network structures. *DESIDO Journal of Library & Information Technology*, *31*, 340–347.
Ruiz-Castillo, J., & Costas, R. (2014). *The skewness of scientific productivity. Working paper 14-02*. Universidad Carlos III. http://hdl.handle.net/10016/18286
Ruiz-Castillo, J., & Waltman, L. (2014, March). *Field-normalized citation impact indicators using algorithmically constructed classification systems of science. Working paper 14-03*. Universidad Carlos III. http://hdl.handle.net/10016/18385
Schneider, J. W., & Costas, R. (2013). *Bibliometric analyses of publications from Centres of Excellence Funded by the Danish National Research Foundation, Report to the Danish Ministry of Science, Innovation and Higher Education Danish Centre for Studies in Research and Research Policy*. Denmark: Department of Political Science and Government, Aarhus University. http://dg.dk/filer/Publikationer/Evaluering2013/Appendiks%205_bibliometrisk_report_03122013.pdf
Schubert, A., Glänzel, W., & Braun, T. (1987). A new methodology for ranking scientific institutions. *Scientometrics*, *12*, 267–292.
Seglen, P. (1992). The skewness of science. *Journal of the American Society for Information Science*, *43*, 628–638.
Tijssen, R., Hollanders, H., & van Steen, J. (2010). *Wetenschaps en Technologie Indicatoren 2010*. Neederlands Observatorium Wetenschap en Technologie (NOWT).
Wagner-Döbler, R. (1995). Where has the cumulative advantage gone? Some observations about the frequency distribution of scientific productivity, of duration of scientific participation, and of speed of publication. *Scientometrics*, *32*, 123–132.
Wagner-Döbler, R., & Berg, J. (1995). The dependence of Lotka's law on the selection of time periods in the development of scientific areas and authors. *Journal of Documentation*, *51*, 28–43.
Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*, 2378–2392.
Waltman, L., & Van Eck, N. J. (2013a). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison. *Scientometrics*, *96*, 699–716.
Waltman, L., & Van Eck, N. J. (2013b). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, *7*, 833–849.
Waltman, L., Van Eck, N. J., & Van Raan, A. F. J. (2012). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, *63*, 72–77.