Scientometrics (2021) 126:1495–1525
https://doi.org/10.1007/s11192-020-03801-1

The lack of meaningful boundary differences between journal impact factor quartiles undermines their independent use in research evaluation

Gabriel-Alexandru Vîiu[1] · **Mihai Păunescu**[1]

Received: 10 July 2020 / Published online: 4 January 2021

Abstract
Journal impact factor (JIF) quartiles are often used as a convenient means of conducting research evaluation, abstracting the underlying JIF values. We highlight and investigate an intrinsic problem associated with this approach: the differences between quartile boundary JIF values are usually very small and often so small that journals in different quartiles cannot be considered meaningfully different with respect to impact. By systematically investigating JIF values in recent editions of the Journal Citation Reports (JCR) we determine it is typical to see between 10 and 30% poorly differentiated journals in the JCR categories. Social sciences are more affected than science categories. However, this global result conceals important variation and we also provide a detailed account of poor quartile boundary differentiation by constructing in-depth local quartile similarity profiles for each JCR category. Further systematic analyses show that poor quartile boundary differentiation tends to follow poor overall differentiation which naturally varies by field. In addition, in most categories the journals that experience a quartile shift are the same journals that are poorly differentiated. Our work provides *sui generis* documentation of the continuing phenomenon of impact factor inflation and also explains and reinforces some recent findings on the ranking stability of journals and on the JIF-based comparison of papers. Conceptually there is a fundamental problem in the fact that JIF quartile classes artificially magnify underlying differences that can be insignificant. We in fact argue that the singular use of JIF quartiles is a second order ecological fallacy. We recommend the abandonment of the quartiles reification as an independent method for the research assessment of individual scholars.

Keywords Journal impact factor (JIF) · JIF quartiles · Journal citation reports (JCR) · JCR subject categories · Meaningful differences · Local quartile similarity

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11192-020-03801-1) contains supplementary material, which is available to authorized users.

✉ Gabriel-Alexandru Vîiu
gabriel.viiu@yahoo.com

[1] National University of Political Studies and Public Administration, Bucharest, Romania

Introduction

Continuing the longstanding tradition initiated by the Institute for Scientific Information in the 1970s each year Clarivate Analytics publishes the annual Journal Citation Reports (JCR). Their stated mission is that of offering "a thorough, publisher-independent, multifaceted view of journal performance, reflecting the world's highest-quality scientific and scholarly literature" (Collier 2019). The main attraction for journal editors, scholars, funding bodies, administrators and others is offered by the publication of a specific journal metric within JCR, namely the journal impact factor (JIF). The JIF was initially promoted by Eugene Garfield (Garfield 1972, 1990) as a way of assessing the impact of different journals independent of their size, i.e. the number of published articles. Because it approximates the mathematical mechanics of an arithmetic mean and therefore has a deceptive simplicity the JIF has become popular not only as a retrospective but also as an a priori impact assessment tool. Problematically, it is used to assess the distinct articles published in a journal and the merit of individual scholars. Owing to the field-specific characteristics of citation distributions and to well-documented technical limitations of the JIF itself (Archambault and Larivière 2009; Seglen 1997; Vanclay 2012) such practices have been repeatedly and forcefully contested. They nonetheless continue unabated in a variety on institutional settings and policy contexts involving research evaluation. As remarked several year ago a certain "impact factor style of thinking" (Fernández-Ríos and Rodríguez-Díaz 2014) seems to have cemented among researchers and organizational actors. This makes discussion of the JIF and JIF-related topics ever current subjects of debate and scrutiny, especially in contexts in which this metric or criteria directly derived from it play an essential evaluative role for individual scholars.

Within the "JIF-dependent managerialism of modern science" (Curry 2018) the exact use of JIFs in research evaluation can take several forms. Among them is the use of hierarchical impact factor classes which pool together journals based on the ranking constructed atop individual JIF values. Having more articles in the top JIF classes is then equated with a higher performance. In the annual JCR in which JIFs are presented a distinctive classification is offered by default in which journals are assigned to one of four JIF quartiles—Q1, Q2, Q3 and Q4—built from the underlying JIF values. Q1 journals (where the highest 25% JIF journals are found) are considered more impactful than Q2 journals, those in Q2 more impactful than those in Q3 and those in Q3 more impactful than those in Q4 (which have the lowest JIFs). However, a journal's JIF quartile has a fundamentally relative meaning: it depends on the journal's assignement to the 236 specific JCR categories within which JIF values are reported. A journal may be assigned to multiple categories and may therefore belong to different JIF quartiles in its different host categories.

The quartile classification has been used in the Spanish "sexenio" in which researchers can submit publications every six years to obtain salary increases (Rafols and Robinson-Garcia 2016). In China papers in journals that belong to the first two JIF quartiles are required in some universities to secure academic tenure (Shu et al. 2020). In addition, a mechanism based on a variation of the original JCR quartiles (prescribed by the Chinese Academy of Science) is reported to be the most frequent basis on which Chinese universities allocate direct monetary incentives to their staff in order to increase research output (Quan et al. 2017).[1] The JIF quartile system also plays a distinct role in some Romanian

[1] Since 2018 China has surpassed the United States of America and is the top ranking country with respect to research articles and reviews indexed in the Science Citation Index Expanded (Liu 2020). The above-mentioned policies have no doubt contributed to this performance. However, Zhu (2020) has recently reported that 2020 is a year of major reform for the evaluation of research in China: recent policy docu-

policy practices concerning the funding of universities and within a national program for rewarding scholars for research outputs.

The preceding examples are meant to highlight a key policy premise of the current paper: the JIF quartile classification system is de facto employed—independent from other indicators or contextual information—as an officially sanctioned appraisal mechanism for individual researchers. A second, equally important but more technical premise that underpins the present investigation is the following: JIFs are reported in the JCR as numerical values with a precision of three decimal places but the transition from the actual *numeric* JIFs to the *categorical* JIF quartiles essentially obliterates the scale and meaning of the initial numeric values. It also obscures the magnitude of the underlying differences between JIFs in different classes for anyone without access to (or interest in) the complete JCR data. The key issue is that since differences between individual JIF values may be very small one may also expect the differences between the resulting quartile classes to also be quite small. Of particular interest are the threshold values that separate the lower bound of an upper quartile (say Q1) from the upper bound of a lower quartile (Q2; likewise for Q2 and Q3, Q3 and Q4). It is not unreasonable to expect that the differences between JIF quartile thresholds are small enough to make them insignificant for the purpose of meaningful evaluation and therefore call into question their use as singular policy drivers.

As a simple example consider that a JIF value of 3.273 conveys essentially the same substantive information as a JIF of 3.272. However, a "Q1" label derived from the first of these two JIF values conveys an altogether different message from a "Q2" label that may be derived from the second value. The simple act of conversion to quartile classes fundamentally alters the perception regarding impact by artificially magnifying an insignificant difference. It is not difficult to see that this problem generalizes to other journals that may have very similar JIFs. The question remains whether empirically very small differences are indeed to be found at the quartile class boundaries and whether this is a common or rather an infrequent occurrence. In general, the lack of meaningful boundary differences is a very common and intuitive issue for categorical variables derived from continuous numeric variables. We therefore have reason to hypothesize a non-negligible incidence of small differences between JIF quartile classes. However, the exact scale of the phenomenon and its concrete manifestation across scientific fields remain open empirical questions which so far have not been systematically addressed. We believe detailed knowledge of non-meaningful boundary differences in the context of JIF quartiles is especially relevant for their concrete policy use, especially in the evaluation of individual scholars for hiring, promotion or pecuniary rewards. At a time when "thinking with indicators" (Müller and de Rijcke 2017) is increasingly affecting the process of scientific research we believe the nuances and limitations of the indicators deserve more scrutiny.

The aim of this paper is to systematically investigate whether very small journal impact factor differences that may be artificially enlarged through quartile aggregation occur in practice and to further determine if such occurrences can be a cause for concern in policy contexts. Data from 4 consecutive editions of the Journal Citation Reports—2015 through 2018, Science Citation Index Expanded and Social Sciences Citation Index—are used to this end. The entire set of 236 JCR subject categories delineating scientific fields is investigated and the differences between the JIFs of the journals that sit on the thresholds of

Footnote 1 (continued)
ments jointly issued by the Ministry of Education and the Ministry of Science and Technology explicitly de-emphasize Science Citation Index publications.

quartile classes are examined. Based on a literature-informed normative definition of *minimally meaningful* JIF differences we determine the share of poorly differentiated journals in each category. We also propose a quartile similarity measure to specifically explore the homogeneity of JIFs in the regions bordering consecutive quartile classes. We find that as a rule boundary differences between consecutive JIF quartile classes are very small and, in fact, so small that they cannot be considered minimally meaningful. Furthermore, we show that it is typical to see between 10 and 30% poorly differentiated journals in the JCR categories with social sciences usually more affected than science categories. We also trace this issue to the characteristic JIF dispersion in the JCR categories which we quantify exhaustively by considering all possible pairs of within-field JIF differences. Finally, we show that temporal quartile instability is closely linked with poor boundary differentiation for most JCR categories. Our findings caution against the independent use of quartiles for purposes of research evaluation and we dispute the singular use of quartiles especially at the micro level of authors. Adding to the empirical material we argue that from a conceptual standpoint the singular use of the JIF quartile system is an unfortunate second order ecological fallacy. It compounds, rather than addresses the first order ecological fallacy of ascribing to a paper the impact merits of its publishing journal. The latter problem has already been denounced (e.g.: Leydesdorff et al. 2016) but this has not prevented many from succumbing to the former.

Our paper is structured as follows: we first offer an overview of the mechanics involved in the calculation of JIFs and the derived JIF quartile classes in "The impact factor and the quartile classification of journals". The "Methods and data for assessing the differences between the JIF quartile boundaries" section presents the approach used to investigate the problem of quartile boundary differentiation. The "Results" section presents our empirical findings and the "Discussion and conclusions" section examines them in relation to previous literature.

The impact factor and the quartile classification of journals

Since there is an abundant literature on JIFs we stress that this paper is not concerned with reenacting an already decades old broad dispute in which the main lines of contention have by and large been clearly drawn and repeatedly addressed. For a recent and comprehensive account the reader is referred to Larivière and Sugimoto (2019). Our aim is to investigate the very specific technical aspect of converting numerical JIF values to the categorical quartile representation and discuss the limitations of this process. For a self-contained reading we find it necessary to review only some essential points regarding the calculation of JIFs. This will help contextualize the discussion of the JIF quartile classes and provide a minimal requisite framework for our methods, data and results.

The calculation of the journal impact factor

According to the official definition the JIF represents (Clarivate Analytics 2018, p. 8):

> all citations to the journal in the current JCR year to items published in the previous two years, divided by the total number of scholarly items (these comprise articles, reviews, and proceedings papers) published in the journal in the previous two years.

One may note from the above definition that on the scaffold of a simple ratio between two quantities three distinct features immediately invite analytical scrutiny: (1) the use of *all* citations; (2) a constrained *two-year* window for the items whose citations are counted; (3) a potential discrepancy between "items published" in the *numerator* and "scholarly items" that in the *denominator* are restricted to very specific document types. These features are addressed in turn in the following paragraphs.

First, in the calculation of JIFs not all items are taken into account, but only those defined as "citable" (articles, reviews, and proceedings papers as mentioned above). This can affect journals very differently based on what type of content they publish. Historically, there is a well-documented and contested numerator—denominator asymmetry which is reported (Archambault and Larivière 2009) to be one of the most commonly mentioned technical limitations of the JIF. In defense of this asymmetry it has been argued that the citable items account for most of the citations accumulated by the publications in a journal. As an example, in a study of 1.1 million items published in 2000 and assessed against an eight-year citation window (McVeigh and Mann 2009) it was reported that 97.6% of citations were accounted for by the citable items. Nonetheless, variation for different journals can still be expected.

The second fundamental feature of the standard JIF which must be emphasized is that the two-year citation window is nothing more than a convention, albeit one with serious implicit consequences. A deficiency that may not be immediately perceivable in this convention was recognized early on by its proponent: by the choice for the very short citation window the JIF "tends to favor research areas that more heavily cite recent research published in the last 2 years" (Garfield 1990, p. 188). This creates several types of problems, chief among which is the fact that the JIF is "biased in favor of journals that have a rapid rather than a prolonged impact" (Vanclay 2009, p. 3). In recognition of the limitations associated with a two-year citation window and of its potential to put some journals and research fields at a disadvantage a five-year journal impact factor has also become a standard offering in the annual JCR. However the five-year JIF has not become a flagship indicator to replace the default two-year JIF[2] and neither have the more complex metrics that seek to appraise the influence of a journal.

The third fundamental feature inherent in the very design of the JIF is its use of *all* the citations received by the publications of a journal. This effectively allows self-citations (citations from a journal article to the other articles published in the same journal) to be fully factored into the calculation of this metric. It has been argued that especially against the background of the short two-year citation window the standard JIF "will be highly influenced by self-citations" (Aksnes 2003, p. 242). This opens doors and creates stimuli for manipulating this metric during the editorial process, including outright coercion of authors to include specific citations to the journal's publications (Wilhite and Fong 2012). A phenomenon of "impact factor biased journal self-citations" in which a journal's self-citations are increasingly directed specifically towards publications from the most recent 2 years (which correspond with the JIF citation window) has been documented by Chorus and Waltman (2016). The JCR also offers a JIF which excludes self-citations but this is not the default used in the quartile classification system.

[2] It is worth noting in passing that Elsevier launched in late 2016 a direct rival to the JIF in the form of the "CiteScore" (Zijlstra and McCullough 2016) together with an accompanying set of CiteScore metrics. The use of a compromise three-year citation window (recently changed to four years) was the main argument offered in support of the alternative metric's "robust approach".

In addition to the preceding three features that are essentially observable in the JIF's definition a further aspect which is not must also be noted. Because in itself a JIF is nothing more than an informational shortcut it is only its placement in a comparative context that infuses it with substantive meaning and any added value. The relevant comparative context for a JIF is created by the scientific field in which the journal it belongs to is published. Technically JIFs are reported in the JCR based on the assignment of journals to a granular categorization which currently includes 236 fields known as "subject categories". These cover sciences, i.e. the Science Citation Index Expanded (SCIE) with its 178 categories, and social sciences, i.e. the Social Sciences Citation Index (SSCI) with its 58 categories. Within each of these subject categories the affiliated journals are ranked—by default—according to their JIF thereby conveying a clear image regarding the relative impact differences between them. Unfortunately, the subject classification itself is far from perfect as a representation of scientific fields: it was developed foremost for information retrieval rather than for scientific impact assessment (Leydesdorff and Bornmann 2016), it lacks transparency (Wang and Waltman 2016) and it is known to suffer from "erroneous lumping of unconnected journals into a single category" (Milojević 2020, p. 194). In addition, many journals are assigned to more than one category because it is not always possible to uniquely map their content to a single scientific field.

Without delving into additional details it is obvious that the JIF features discussed above present strong limitations for taking very small differences between 2 JIFs as being substantively meaningful. Nonetheless, as stated in the introduction the JIF is traditionally and immovably reported since its very introduction in the '70 s to a degree of precision involving three decimals. What is worth particular emphasis in the context of the present paper is the fact that this decision seems to have been motivated purely by practical reasoning, rather than from intrinsic belief in the precision of the JIF. As recognized by Garfield in an article for the *Journal of the American Medical Association*, "the precision of impact factors is questionable, but reporting to 3 decimal places reduces the number of journals with the identical impact rank" (Garfield 2006, p. 93). While it may indeed present this tie-breaking advantage the use of the three decimal places has been repeatedly criticized and rejected as unsound. Summarizing an at the time already extensive literature Vanclay (2012, p. 220) listed "data errors, system faults, sampling deficiencies and statistical shortcomings" as elements casting doubt on the reproducibility of the JIF and on the three decimals convention used in its publication. In a similar vein reporting the JIF with three decimals has been offered as a prime example of unwarranted "false precision" in the scientometric community where it has been argued that "given the conceptual ambiguity and random variability of citation counts, it makes no sense to distinguish between journals on the basis of very small impact factor differences" (Hicks et al. 2015, p. 431). Vanclay as well as Hicks and colleagues specifically support the use of a single digit JIF.

The only explicit rebuttal that seems to have been offered to the JIF precision problem is the following (Pudovkin and Garfield 2012, p. 410):

> In presenting fractions in decimal form one needs (in order not to lose information) to retain as many digits after the decimal point as there are digits in the denominator. The denominator in this case is the number of citable items in the 2-year set of a journal. For many journals this number is hundreds, sometimes thousands. So, three digits are fully justified. And yes, this eliminates many ties.

This rebuttal is unconvincing since from the aforementioned questionable precision already conceded for JIFs it does not follow that the information being preserved in the conversion from fraction to decimals is rigorous enough to merit preservation to such a

high degree. To the contrary, a rounded approach seems more justified. Nonetheless, based on a simple scenario of adding or removing a single citation and determining the effect on JIFs it was argued that for about two thirds of JCR journals the 3 decimal convention "can be considered mostly adequate" (Campanario 2014, p. 292).

The journal impact factor quartile classes

This section provides a short focused discussion of the specific topic of JIF quartile classes and presents the explicit questions to be investigated in the empirical part.

Anyone that has a fair acquaintance with statistics will in general immediately associate the idea of quartiles with the boxplot. In this graphical representation of numerical data the distribution of values is segmented in 4 distinct parts. Information on the spread of the observed cases is conveyed by marking the minimum, the 1st quartile, 2nd quartile (i.e. the median), 3rd quartile and the maximum. As a display of numeric JIF values from a given subject category this type of plot is a simple but effective choice. The suggestion that the various JCR journal metrics, including the impact factor, could be more usefully understood if presented within *groups* derived with the aid of this now standard statistical device seems to have first been proposed more than 2 decades ago by Magri and Solari (1996).[3] The idea of comparing individual researchers by using a "%Q1 indicator [defined as the] ratio of publications that a researcher has published in the most influential journals" (Bornmann and Marx 2014, p. 496) has also been mentioned as a way of comparing researchers in the natural and life sciences. More recently publications in the Q1 journals have been used to compare countries (Madhan et al. 2020) and a proposition of a quartile class weighted JIF (Adigozalova 2019) has been put forward to capture the distinct influence of citing sources.

The current implementation of JIF quartiles in the JCR[4] is not strictly based on the actual JIF values, but rather on the *percentile ranks* that are associated with this metric. The JIF percentile rank is a ratio between the actual rank given within a subject category by the JIF and the number of journals from that category. Based on its JIF percentile rank (PR) a journal is attributed to Q1 if its PR is located in the interval (0.0, 0.25], in Q2 if its PR is in the (0.25, 0.5] interval, in Q3 if the PR is in (0.5, 0.75] and in Q4 if its PR is above 0.75. Ideally, assuming for example 100 journals in a field and no ties in JIF values there would be exactly 25 journals in each quartile. Recall however that even the 3 digit decimal convention apparently only *reduces* the number of journals with identical rank so in practice a JIF quartile may not contain an exact quarter of all the JIF values. Carefully browsing through JIF rankings shows that when ties occur between journals the minimum rank value (and therefore percentile rank) is attributed to each. This means that if, for example, the journals on the 50th and 51st position would have an identical JIF they would both be placed in Q2. However, if three journals on the 26th, 27th and 28th position would have an identical JIF—even if only marginally smaller than a journal on the 25th place in Q1– they too would be placed in Q2.

[3] From the current JCR scope notes for *category box plot* it seems this specific article is the inspiration for the current implementation of the JIF quartile classes as a way to compare journal performance across categories.

[4] The account provided in this paragraph follows the description provided in the InCites Indicators Handbook (Clarivate Analytics 2018, p. 10).

In view of the above it is reasonable to be at least moderately prudent if not altogether distrustful when presented *only* with the information that a journal is in Q1, Q2, Q3 or Q4, or that an author has published a paper in a journal from a specific quartile. The well-established features of the JIF offer some theoretical and conceptual caution against the indiscriminate use of quartiles. However, it is the specific possibility of ties and the high degree of precision with which JIFs are adamantly reported that invite a detailed empirical analysis of the magnitude of the differences that actually separate the JIF quartile classes. At least three specific and interwoven issues seem of interest and—considering each distinct JCR subject category in a given year—they may be summarized in three interconnected but technically distinct exploratory research questions:

RQ1: What is the empirical magnitude of the JIF differences separating the quartile classes?
RQ2: Are the differences between the limiting values of the quartile classes—as a rule—at least minimally meaningful?
RQ3: Assuming some normative minimally meaningful difference among impact factors, how many journals—if any—lie within the spaces that preclude meaningful distinction between the quartile classes?

The scope of these questions is further explained and formally stated in the following section.

Methods and data for assessing the differences between the JIF quartile boundaries

Our inquiry encompasses the level of JCR subject categories. To answer RQ1 it is necessary to identify, within each subject category, the JIF limiting quartile values and then determine the distances between them. Some basic mathematical formal modeling will help make the idea explicit and also contribute to presenting the treatment of RQ2 and RQ3. First, consider the ordered pair (H, L) as representing any of the 3 possible pairs of consecutive JIF quartile classes (Q1, Q2), (Q2, Q3) or (Q3, Q4), determined by the percentile ranking method previously described. H is the set of JIF values from the higher class and L is the set of values from the lower JIF class. Within each quartile class the JIFs are sorted in descending order. Denoting the lowest JIF value in a quartile with λ and with υ the highest JIF value in a quartile, the results for the first research question will simply consist in determining, for each of the three possible (H, L) configurations, the empirical differences of the form $\lambda^H - \upsilon^L$.

To answer RQ2 and RQ3 the *minimally meaningful difference*, which will be denoted with δ, must be defined. Contingent on the specification of δ the answer to RQ2 consists in determining for each (H, L) pair whether the inequality

$$\lambda^H - \upsilon^L > \delta \tag{1}$$

is satisfied or not. In practice δ will of course be a specific constant that is unfortunately subject to the idiosyncrasy of choice. As an additional challenge one might also argue that meaningful differences should be field-dependent and that each JCR category is amenable to its distinct δ specification. To address these issues we deliberately opt to answer our research questions in the restricted framework of a *minimally meaningful difference*

which has the advantage of cross-field applicability. We hold that such a minimal absolute standard can be derived from previous works that strongly recommend—irrespective of the scientific field of a journal—using only one decimal for JIF values (Hicks et al. 2015; Vanclay 2012). Based on this unambiguous recommendation a lower bound for δ demands a value of *at least* 0.1. We therefore retain this value as a *minimum minimorum* δ for all JCR categories. Higher δ values might be appropriate, especially for fields where high JIF values are common, but it is evident that even for these a *minimally* meaningful difference should not strive to be very large. On the other hand, to offer a clear answer to RQ2 and RQ3 we believe it is essential to specify a single, a priori benchmark rather than a *post-hoc* plurality of field-specific benchmarks. The selection of the latter would be far from obvious and uncontroversial. In this paper 0.1 will be retained as the default benchmark for minimally meaningful differences due to the advantage of cross-field applicability. However, values of 0.2 and 0.3 will also be considered in the empirical exploration to offer a sensitivity analysis of our results. For each of the three δ variations the answer to RQ2 will either be positive (indicating the existence of the minimally meaningful difference between two quartiles) or negative (indicating the absence of this difference).

Finally, for RQ3 the idea is to determine the number of journals for which the JIFs are found within a distance of δ around the limiting values that separate any pair of adjacent quartiles. There is, however, no compelling reason to adopt a one-way appraisal. From the perspective of L, the interval of interest is $[\upsilon^L, \upsilon^L + \delta]$; note that if the answer to the second research question above is negative this interval will extend upward into H at least to the level of λ^H but it might possibly encompass higher values. Analogously, from the perspective of quartile H the interval of interest is $[\lambda^H, \lambda^H — \delta]$; if the answer to the second research question above is negative this second interval will extend downward into L at least to the level of υ^L but possibly even lower. Condensing the two conditions and introducing the notation φ to designate an arbitrary JIF value belonging to an (H, L) pair under analysis the formal answer to RQ3 is expressible as identifying the set(s) of journals

$$F = \{\varphi \in (H \cup L) | \lambda^H - \delta \leq \varphi \leq \upsilon^L + \delta\} \tag{2}$$

If F is empty (which will coincide with a positive answer to RQ2) then 2 quartile classes may be said to be meaningfully different, recalling of course the exact specification of the δ parameter. Conversely, if F is not empty, more values in the set will indicate an ever weaker differentiation between consecutive JIF quartile classes with respect to δ. Since there are three possible pairs of consecutive quartile classes in each subject category there will be three distinct F sets. To disambiguate we use the notation F(I) to refer to the F set corresponding to the (Q1, Q2) pair, F(II) for the one corresponding to (Q2, Q3) and F(III) for the (Q3, Q4) pair.

As a general answer to RQ3 we aim to determine the share of the journals that belong to the F(I), F(II) and F(III) sets within each JCR category. We also aim to quantify the level of differentiation or homogeneity in each of the three δ spaces around the quartile boundaries for a given subject category. To this end we define a more specialized metric labelled the indicator of *local quartile similarity* (LQS). For each (H, L) pair

$$\text{LQS} = \frac{|F|}{|H \cup L|} * 100 \tag{3}$$

There are 3 advantages to using this percentage indicator. First, it allows *within* subject category comparisons: it becomes possible to assess if, for example, differences between Q1 and Q2 are more pronounced than those between Q2 and Q3. The second advantage of

Table 1 Hypothetical JIF values, ranking and quartile classes in a fictitious subject category, highlighting poorly differentiated journals and their assignment to specific F sets under three δ variations

Journal	JIF	Journal rank	Percentile rank	JIF quartile class	F set journal with $\delta = 0.1$	F set journal with $\delta = 0.2$	F set journal with $\delta = 0.3$
Journal A	7.364	1	0.038	Q1			
Journal B	7.175	2	0.077	Q1			
Journal C	3.501	3	0.115	Q1			
Journal D	2.535	4	0.154	Q1			
Journal E	2.221	5	0.192	Q1		F(I)	F(I)
Journal F	2.057	6	0.231	Q1	F(I)	F(I)	F(I)
Journal G	2.029	7	0.269	Q2	F(I)	F(I)	F(I)
Journal H	1.459	8	0.308	Q2			
Journal I	1.448	9	0.346	Q2			
Journal J	1.399	10	0.385	Q2			
Journal K	1.189	11	0.423	Q2			
Journal L	1.030	12	0.462	Q2		F(II)	F(II)
Journal M	1.007	13	0.500	Q2		F(II)	F(II)
Journal N	0.854	14	0.538	Q3		F(II)	F(II)
Journal O	0.815	15	0.577	Q3		F(II)	F(II)
Journal P	0.793	16	0.615	Q3			F(II) & F(III)
Journal Q	0.608	17	0.654	Q3	F(III)	F(III)	F(III)
Journal R	0.602	18	0.692	Q3	F(III)	F(III)	F(III)
Journal S	0.582	19	0.731	Q3	F(III)	F(III)	F(III)
Journal T	0.509	20	0.769	Q4	F(III)	F(III)	F(III)
Journal U	0.441	21	0.808	Q4		F(III)	F(III)
Journal V	0.368	22	0.846	Q4			F(III)
Journal W	0.292	23	0.885	Q4			F(III)
Journal X	0.279	24	0.923	Q4			
Journal Y	0.241	25	0.962	Q4			
Journal Z	0.152	26	1.000	Q4			

the indicator is that it accounts for the different sizes of the subject categories themselves because the denominator in Eq. 3 will always be approximately ½ of the complete number of journals in a subject category. The indicator can therefore be used to make comparisons *between* the categories and *across* time. For example, a higher LQS in a subject category X than in subject category Y would indicate that for X the problem of poorly differentiated journals concentrated around the quartile boundary is of greater concern for the specific (H, L) pair analyzed. Finally, LQS has an alternative appealing interpretation. It provides the empirical probability that a randomly selected journal from a specific (H, L) pair will occupy the δ space around the quartile thresholds that marks a lack of minimally meaningful differentiation. The higher such a probability, the more difficult it is to admit the existence of meaningful differences between the 2 quartile classes in question. Similar to the notation used to disambiguate F sets, in the subsequent sections LQS(I) will represent the value of the LQS indicator for the (Q1, Q2) pair, LQS(II) the value for the pair (Q2, Q3) and LQS(III) the value of the indicator for the (Q3, Q4) pair.

We illustrate in Table 1 the fundamental aspects of our investigation with hypothetical JIF data[5] for a fictitious subject category with 26 journals labelled A through Z:

(1) The pairwise absolute differences between limiting JIF quartile classes are 0.028 for the difference Q1–Q2 that separates Journals F and G, 0.153 for Q2–Q3 (separating journals M and N), 0.073 for Q3–Q4 (separating Journals S and T). The mean quartile difference for this subject category would be 0.085.
(2) With regard to the F sets comprising journals with poorly differentiated JIFs, considering $\delta = 0.1$ we have two journals in F(I), namely F and G, no journals in F(II) and four journals in F(III), namely Q, R, S and T. In total, 6 of the 26 journals are poorly differentiated, meaning a share of about 23%. It is obvious however that there are sharp differences between the levels of homogeneity at the 3 quartile class boundaries that are not captured by the total percent share of affected journals. Resorting to our LQS indicator we have LQS(I) = 15.38% (2/13), LQS(II) = 0% (0/13) and LQS(III) = 30.77% (4/13). By rounding each LQS a concise *local quartile similarity profile* for this category can be expressed as 15, 0, 31. Within this category comparing LQS values is of course not dissimilar from comparing the sizes of the three F sets. However, if one were to compare this category with another one with 200 journals, recourse to the size-adjusted LQS values would certainly be more sensible than recourse to the absolute size of F sets and more informative than the global share of poorly differentiated journals.
(3) Imposing a more demanding δ value of 0.2 would increase the number of journals in the F sets as follows: F(I) would comprise three journals, F(II) which was previously void would now have four journals and F(III) would expand to five journals. The total share of poorly differentiated journals would now be 46.15%. LQS(I) would be 23.08%, LQS(II) 30.77% and LQS(III) 38.46%. For the even stricter δ value of 0.3 all indicators increase further and the only additional fact to be emphasized is the position of Journal P which, given the distribution of impact factors, would belong not only in F(II) but also in F(III).

To explore the 3 research questions described above data from the 2015, 2016, 2017 and 2018 editions of the Clarivate Analytics JCR (SCIE and SSCI) are used. The four year period is adequate both for offering a state of the art picture as well as for detecting stable patterns. All distinct JCR subject categories are analyzed: 236 in 2018, 234 in 2015–2017. (For expediency the paragraphs below use the phrase "234 subject categories" regardless of the year.) The next section reports the results that were obtained following the methods described above as well as some additional findings relevant to the study of quartile boundary differentiation.

[5] Hypothetical JIF values were generated in the R language and environment for statistical computing (R Core Team 2020) and follow a standard lognormal distribution; the lognormal distribution is known to adequately approximate various scientometric quantities, especially citations (Brito and Rodríguez-Navarro 2018; Thelwall 2016; Vîiu 2018).

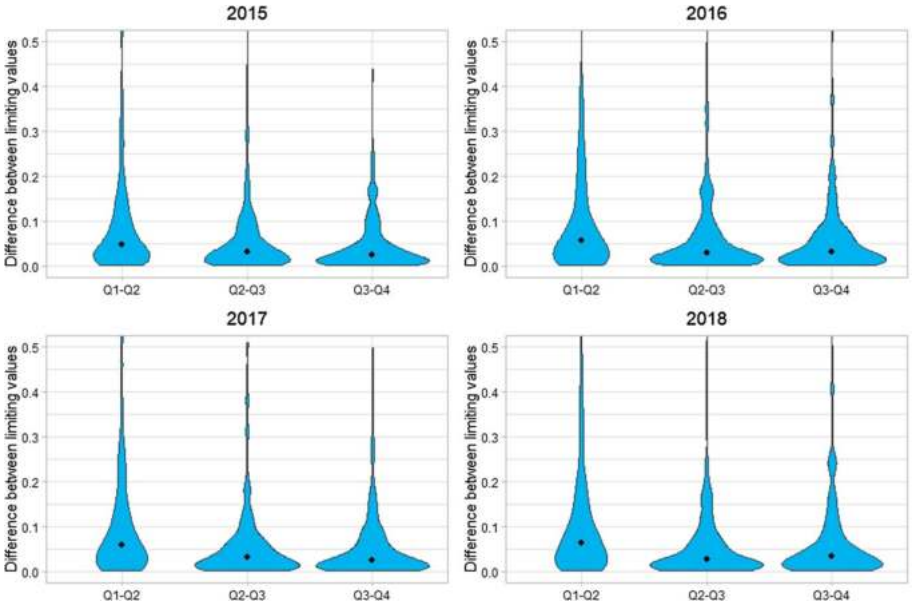

Fig. 1 Violin plots showing the distribution of the 234 subject categories by the absolute differences between the limiting JIF quartile values in 4 recent JCR editions. Y axes are truncated at 0.5 to enhance legibility; inner points mark median values

Results

To begin it is useful to summarize the presentation of our research questions. The first intends to identify the absolute size of pairwise differences between consecutive quartile classes. The second asks if these differences are in general meaningful against a minimal benchmark δ value. The third explores how many journals are affected by poor differentiation at the boundaries of consecutive quartile classes, with poor differentiation conventionally defined by a specific δ.

The scale of the differences between quartile classes

Concise answers to RQ1 and RQ2 are offered in Fig. 1 which provides a high level overview of the absolute differences separating the JIF quartile classes across the 2015–2018 JCR editions. Each distinct shape in the violin plots[6] shows the density distribution of the 234 subject categories with regard to the absolute differences between the limiting values of a specific quartile class pair. Several features immediately stand out within each year: first, for an overwhelming number of subject categories the differences separating consecutive quartile classes are very small. Virtually all pairwise differences in all years are smaller than 0.3 and a substantial majority of differences are smaller than 0.1. A second

[6] See (Hintze and Nelson 1998) for details. All figures in the paper were created in R with the *ggplot2* package (Wickham 2016).

Table 2 Selected percentiles of the distribution of the absolute differences between the limiting JIF quartile values for the 234 JCR subject categories

Quartile pair	Year	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Q1–Q2	2015	0.001	0.006	0.012	0.025	0.032	0.046	0.071	0.098	0.142	0.254	1.366
	2016	0.001	0.006	0.014	0.025	0.036	0.055	0.078	0.113	0.182	0.296	1.534
	2017	0.001	0.007	0.014	0.027	0.041	0.058	0.077	0.119	0.184	0.294	1.231
	2018	0.001	0.006	0.016	0.028	0.043	0.063	0.084	0.115	0.163	0.284	1.194
Q2–Q3	2015	0.001	0.002	0.006	0.012	0.021	0.031	0.040	0.056	0.088	0.134	1.074
	2016	0.001	0.004	0.009	0.013	0.018	0.028	0.038	0.062	0.093	0.169	0.969
	2017	0.001	0.005	0.008	0.012	0.021	0.031	0.042	0.061	0.079	0.137	0.509
	2018	0.001	0.004	0.008	0.013	0.019	0.025	0.042	0.060	0.091	0.160	0.693
Q3–Q4	2015	0.001	0.003	0.006	0.010	0.016	0.024	0.034	0.045	0.079	0.125	0.439
	2016	0.001	0.003	0.008	0.013	0.021	0.030	0.042	0.060	0.080	0.149	0.603
	2017	0.001	0.004	0.007	0.011	0.017	0.024	0.037	0.056	0.083	0.136	0.499
	2018	0.001	0.006	0.009	0.015	0.021	0.034	0.042	0.063	0.100	0.180	0.774

observable feature within each year is that the distribution of Q1–Q2 differences is moderately but notably distinct from the distributions of the Q2–Q3 and Q3–Q4 differences that are more similar. Specifically, Q1–Q2 differences cover a slightly broader range of values and are not as densely concentrated below 0.1. Q2–Q3 and Q3–Q4 differences display a smaller range of values and a much greater concentration in the interval [0, 0.1]. Nonetheless, even Q1–Q2 differences are, as a rule, concentrated below a value of 0.2. Finally, a third essential aspect to note with regard to the violin plots is that only minor variations seem to occur from one year to another. This indicates that the global picture is remarkably stable in time, at least for the 4 years considered in this paper.

A more detailed presentation of the distribution of pairwise class differences and of their variation in time is provided in Table 2. Note (against the purely hypothetical example from the introductory section) that the 0th percentile column which corresponds to the minimum value in the distributions is *always* 0.001. This means that in each year and for each quartile class pair there exists at least one subject category for which the difference between the quartiles is one thousandth of an impact factor of one. Note also (concordant with the graphical representation where it is not readily observable) that one has to look as high as the level of the 70th percentile to observe any instance of a quartile difference of at least 0.1. In addition it is noteworthy that for 80% of the subject categories Q1–Q2 differences display values below 0.2 while for 90% of the subject categories Q2–Q3 and Q3–Q4 differences are lower than 0.2.

Before moving to RQ3 it seems appropriate to delve deeper and investigate the relation between quartile class differences and the size of JCR subject categories. The latter is quite heterogeneous, ranging from as few as 5 journals in the case of *andrology* in 2015 to as many as 363 in the case of *economics* in 2018. It is plausible to hypothesize that very small quartile class differences are especially typical of larger subject categories where more journals can cluster around a specific JIF value. For smaller categories "overcrowding" should be less likely.

Figure 2 explores the relation between *mean quartile differences* (i.e. the averages of the 3 pairwise differences of each category) and journal counts in the context of the 2018 data by employing a local linear regression. The smooth line from the local regression indicates

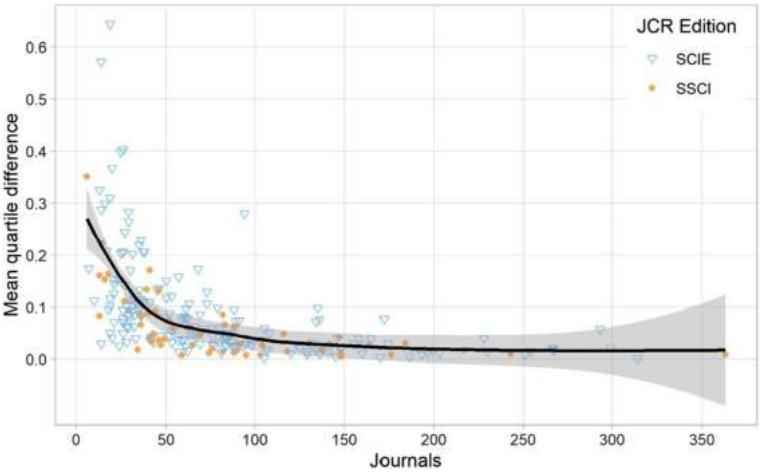

Fig. 2 Local linear regression of mean JIF quartile differences on journal count (2018 data), showing an approximate pointwise 95% confidence envelope and highlighting the science and social science categories. The span parameter in the smoother is 0.5

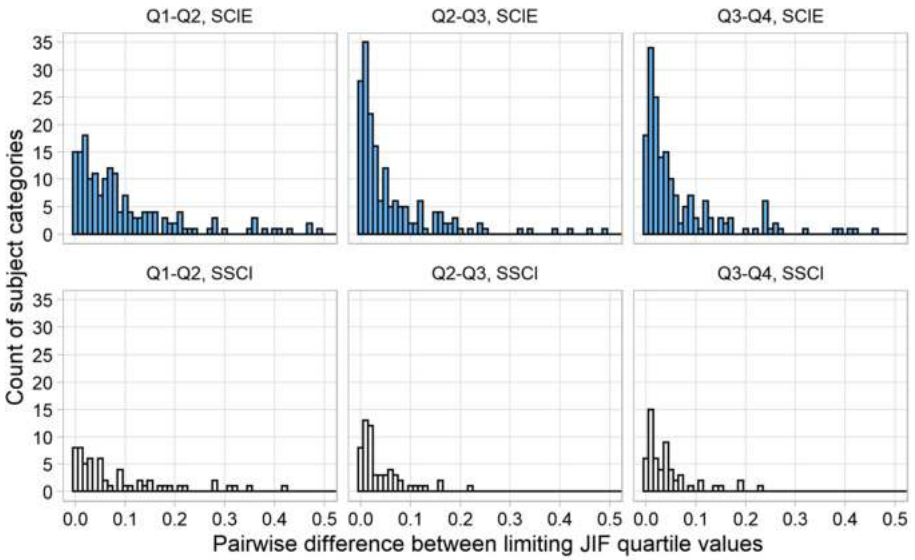

Fig. 3 Histograms of individual JIF quartile class differences within the 2018 JCR subject categories; note the bin width of 0.01 and the fact that the graphs are truncated at 0.5 to enhance legibility

that the mean difference for each subject category tends to decrease with the number of journals, quite steeply in the first part (up to about 30 journals) but then at a continually declining rate. Poor quartile class differentiation is thus not unequivocally traceable to the presence of a greater number of journals: there are many categories which, although having fewer than 50 journals still display a mean quartile difference below 0.1. There is also

no clear pattern when looking to the distinction between science and social science categories; for both the number of journals is a weak predictor for mean quartile differences. However, one additional fact indicated by the figure is that within the social sciences the proportion of categories with poor differentiation seems higher than for the science categories. It turns out that for 2018 49 out of 58 SSCI categories (84%) have a mean quartile difference below 0.1, whereas 127 of 178 (71%) SCIE categories have this property. Essentially similar results are obtained for the data from 2015–2017. Their presentation is thus omitted. For the 2018 data a more detailed depiction of the distribution of *individual pairwise differences* between quartile classes is provided in Fig. 3. Again, results very similar to this figure are obtained for 2015–2017 and they are omitted to avoid redundancy. All pairwise differences for all subject categories in the four year window (as well as F, LQS and other values discussed below) are available in a standard *csv* file as Electronic Supplementary Material 1.

The density of poorly differentiated journals around quartile class boundaries

In answer to RQ3 two aspects are relevant. As a *general* aspect it seems important to determine the overall number of journals that are affected by the poor differentiation around quartile class boundaries irrespective of their exact provenance, i.e. regardless of whether they belong to F(I), F(II) or F(III). Recall each individual F set is defined in accordance with Eq. 2. A second, more *granular* aspect that is relevant for the subject categories is to determine their characteristic indicators of local quartile similarity, the LQS values defined by Eq. 3 which facilitate more in-depth comparison. Furthermore, while the answer to RQ1 is completely independent of idiosyncratic conceptions of what constitutes a meaningful difference and while even the results for RQ2 would not be fundamentally altered by the choice of the δ parameter (as highlighted by the above findings), in the case of RQ3 the parameter specification holds a much greater importance. In addition to our benchmark of $\delta = 0.1$, the weakest scenario of minimally meaningful difference that has cross-field applicability, we also consider values of $\delta = 0.2$ and $\delta = 0.3$ to explore the sensitivity of results to increasingly stringent requirements of differentiation.

Figure 4 addresses the *general* component of RQ3. It provides a comprehensive overview of the percent share of poorly differentiated journals within each category, within each of the 4 years, under each of the 3 δ variations, separately for SCIE and SSCI categories. The results are based on identifying in each category the *unique* number of poorly differentiated journals and determining their share in the full number of journals. For $\delta = 0.1$ in almost all subject categories the unique number of poorly differentiated journals corresponds with summing the F(I), F(II) and F(III) journals because, as a rule, with $\delta = 0.1$ F(I), F(II) and F(III) are disjoined sets. However, in some categories and in some years (specifically in 19 of a total of 938 instances: *history* in 2015–2017, *logic* in 2017 and 2018, etc.) the concentration of JIF values is so great that some journals belong to more than one of the three F sets (similar to the hypothetical Journal P from Table 1). By summing over F(I), F(II) and F(III) they would be double counted and thus yield a slightly enlarged overall number of poorly differentiated journals. To avoid this issue (which increases in intensity with higher δ values) only the *unique* number of poorly differentiated journals was considered.

Restricting the analysis to the benchmark case (δ taken to be 0.1) it is apparent that for the overwhelming number of JCR categories (more than 80%) at least 10% of the journals are affected by poor differentiation. Furthermore, for about half of all the categories at least

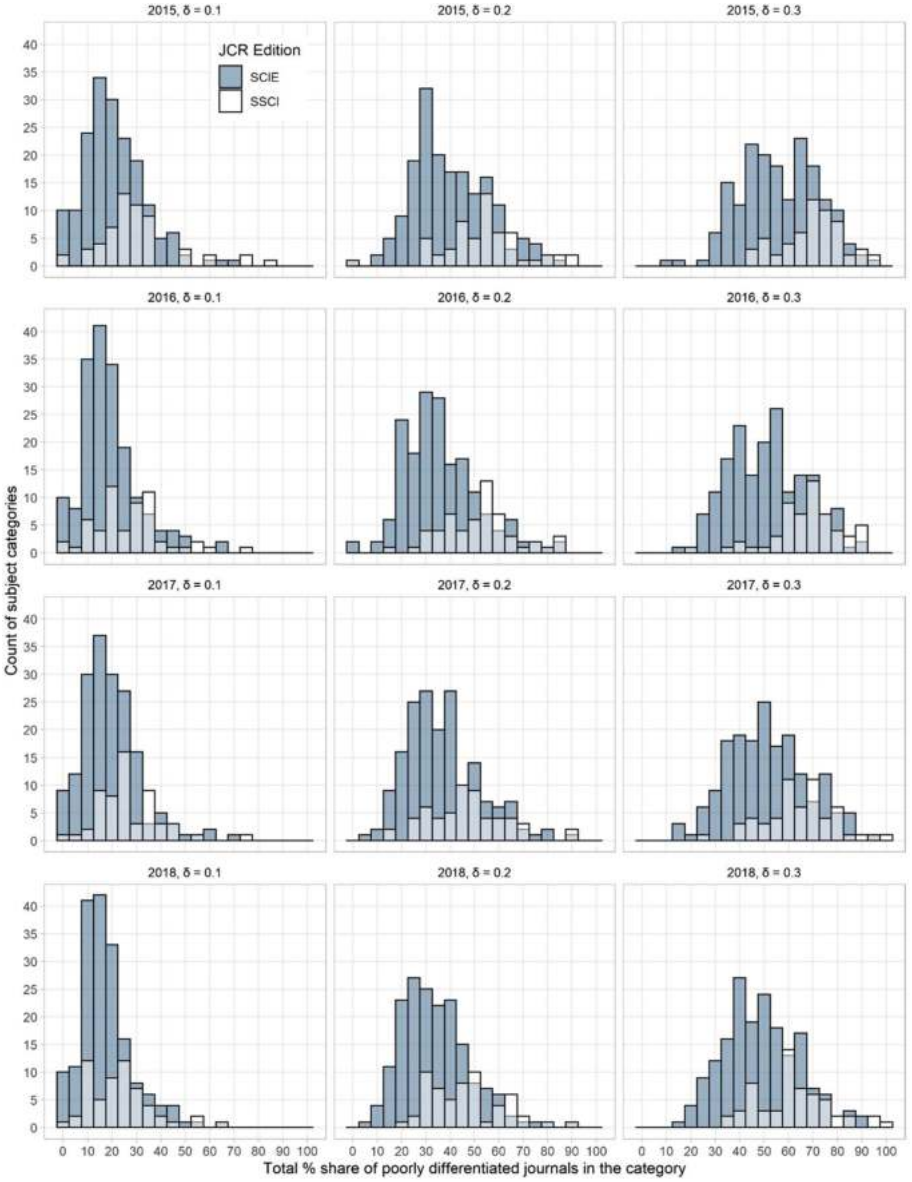

Fig. 4 Distribution of the 234 JCR subject categories viewed with respect to the share of (unique) poorly differentiated journals in the total number of journals. Each subplot displays the results for a specific year and δ variation. SCIE and SSCI categories are shown separately and the bin width used in each overlapping histogram is 5%

20% of the journals are affected by poor differentiation. For about a fifth of all the categories the total share of journals that are poorly differentiated exceeds 30% and extreme cases with more than 50% poorly differentiated journals are also clearly visible. Overall, it is typical to see between 10 and 30% poorly differentiated journals in the JCR categories.

Looking to the distinction between SCIE and SSCI it seems that the latter categories are routinely more affected by the problem than the former. For example, on average across the 4 years about 70% of SSCI categories have at least 20% poorly differentiated journals whereas the same is true for only about 40% of the SCIE categories. About a third of SSCI categories have at least 30% poorly differentiated journals whereas the same is true for only about 15% of the SCIE categories. An interesting albeit subtle phenomenon that merits to be recorded is that overall the problem of poor differentiation seems to be *decreasing discreetly with time*. Moving from 2015 to 2018 we find slightly fewer categories that are affected by the problem of poor differentiation irrespective of the exact value taken as reference. For example, there were 89% SCIE and SSCI categories with at least 10% of journals affected in 2015, but only 82% in 2018. Similarly, there were 39% SCIE and SSCI categories with at least 25% of journals affected in 2015, but only 23% in 2018.

Finally, departing from our benchmark *minimum minimorum* scenario has results that are obvious from the graphical representation. Applying more stringent requirements for meaningful differentiation leads to a dramatic increase in the share of JCR journals affected by homogeneity. With $\delta=0.2$ about 90% of the JCR categories would have at least 20% poorly differentiated journals and about 25% would have at least 50% poorly differentiated journals. With $\delta=0.3$ about 50% of all categories would have at least 50% poorly differentiated journals.

Moving to the *granular* component of RQ3 we now present in more detail the problem of poor JIF differentiation around quartile boundaries by addressing the distribution of the individual LQS indicator values. Figure 5 provides a synopsis of the results obtained for this more discerning indicator not only under the weakest scenario of $\delta=0.1$, but also for $\delta=0.2$ and for $\delta=0.3$. Similarly, Table 3 provides the more detailed account of LQS values for all three scenarios.

Restricting the analysis to the benchmark $\delta=0.1$ and looking to the LQS values obtained for the subject categories across the four years of JCR data at least two essential aspects may be noted. First, as one might expect from the general results regarding the percent share of poorly differentiated journals, there is a non-negligible share of categories for which there are considerably high LQS values. This holds regardless of the exact pair of quartile classes being compared. Second, the problem of poor differentiation around quartile class boundaries is now revealed to be less intense for the (Q1, Q2) pair and more intense for the (Q2, Q3) and (Q3, Q4) pairs. These are shown to be more related, similar to the results already documented for the absolute differences between JIF values. Some selected details in support of these 2 points—evident in Fig. 5 and Table 3—may be explicitly noted. The median LQS(I) value, on average across the four years, is about 7%, while the median LQS(II) and LQS(III) values are about double: 14% and 16%. This means, for example, that for more than half of all JCR categories the share of journals from the (Q2, Q3) pair that belong to the quartile boundary region within which JIF differences are not minimally meaningful is at least 14%. Looking higher to the 80th percentile of LQS values one may note that for 20% of the JCR subject categories the following are attained: LQS(I) values greater than about 15%, LQS(II) values greater than about 23% and LQS(III) values greater than about 26%.

Increasing the requirement of minimal difference also leads to a visible corresponding increase in LQS values. Even for $\delta=0.2$ the median LQS(I) value is about 16% across the four JCR years analyzed and the median LQS(II) and LQS(III) values reach 30%. Isolated instances of maximal LQS values of 100% also emerge. For $\delta=0.3$ the median LQS(I) value clearly exceeds 20% and the median LQS(II) and LQS(III) values move beyond 40%. As an overall remark, the analysis based on the LQS indicator also highlights trends that

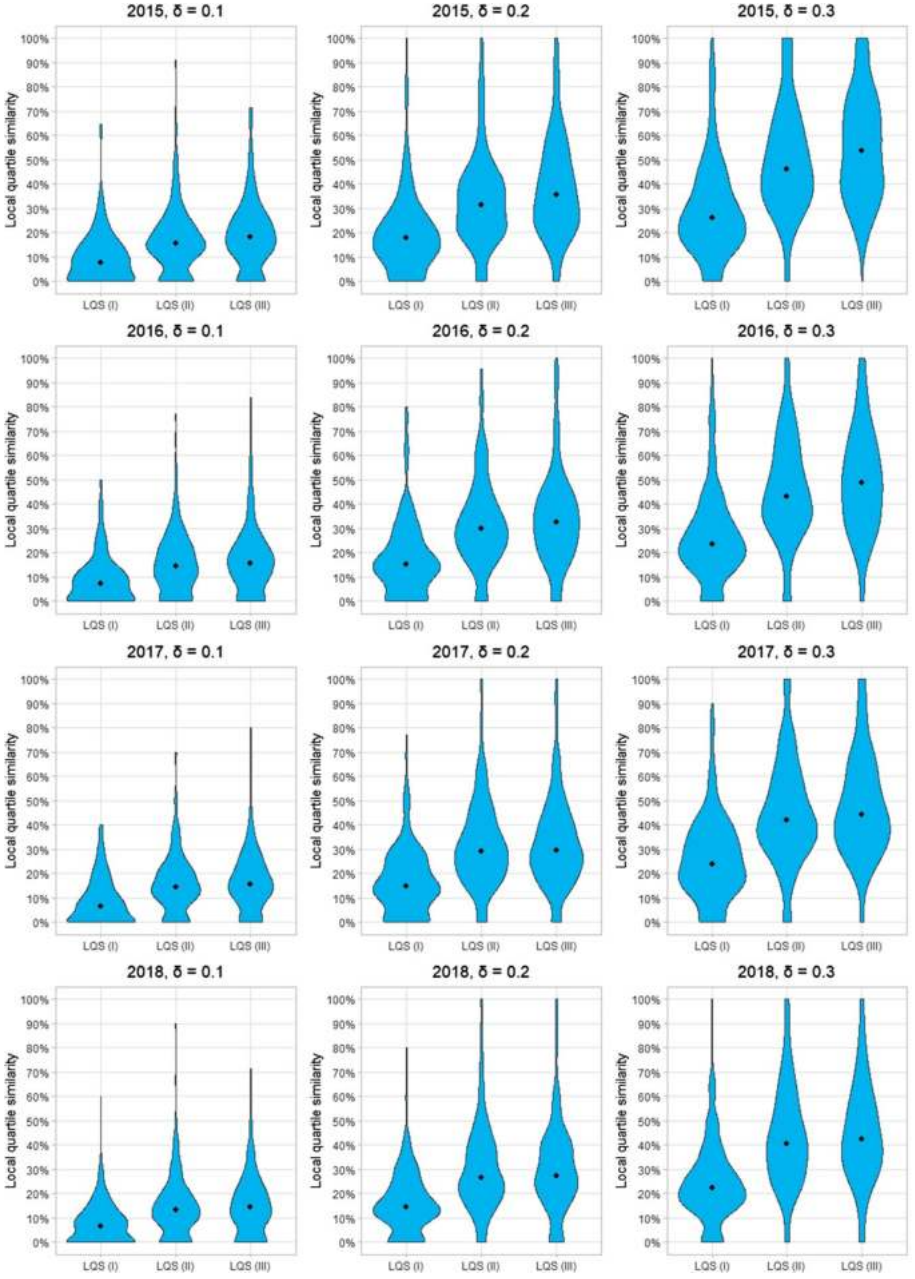

Fig. 5 Distribution of the 234 subject categories by the values of the LQS indicators under three δ variations (points mark median values)

are considerably stable in time. A comprehensive catalogue of the LQS values corresponding to each individual JCR subject category within the last four years is available in Electronic Supplementary Material 1. A detailed visual representation of the LQS profiles for

Table 3 Selected percentiles of the distribution of LQS values for the 234 subject categories under three distinct scenarios of minimally meaningful quartile class differences

δ	LQS	Year	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.1	LQS(I)	2015	0.00	0.00	0.00	1.25	6.19	7.79	10.34	13.57	16.58	23.15	64.71
		2016	0.00	0.00	0.00	0.00	5.03	7.32	9.38	11.77	14.29	23.08	50.00
		2017	0.00	0.00	0.00	0.00	4.56	6.48	8.22	11.64	15.76	22.22	40.00
		2018	0.00	0.00	0.00	0.00	4.76	6.67	8.82	10.82	14.29	18.75	60.00
	LQS(II)	2015	0.00	0.00	6.91	10.70	13.36	15.77	17.70	21.43	24.72	31.75	90.91
		2016	0.00	0.00	4.93	8.69	11.36	14.29	17.50	20.78	25.00	30.61	77.27
		2017	0.00	0.00	6.82	10.26	12.53	14.29	17.52	20.55	23.26	30.88	70.00
		2018	0.00	0.00	5.13	8.79	10.87	13.16	15.57	18.57	22.22	30.48	90.00
	LQS(III)	2015	0.00	0.00	8.75	12.50	15.81	18.18	21.50	24.42	28.57	36.84	71.43
		2016	0.00	0.00	6.43	9.98	13.23	15.62	18.75	20.94	25.00	33.74	84.09
		2017	0.00	0.00	7.76	11.11	13.54	15.70	18.18	22.22	26.46	33.22	80.00
		2018	0.00	0.00	2.99	9.09	11.11	14.29	16.67	20.00	23.53	30.37	71.43
0.2	LQS(I)	2015	0.00	0.00	7.73	11.61	14.29	17.75	20.69	25.00	28.57	40.00	100.00
		2016	0.00	0.00	6.09	9.51	13.19	15.27	17.39	21.48	28.05	35.38	80.00
		2017	0.00	0.00	3.95	10.31	12.50	15.00	17.53	22.63	27.63	31.91	77.27
		2018	0.00	0.00	7.14	10.53	12.50	14.29	17.28	20.83	25.00	32.74	80.00
	LQS(II)	2015	0.00	12.84	18.93	22.84	27.50	31.62	36.44	40.91	45.33	57.14	100.00
		2016	0.00	12.50	18.92	23.08	26.11	30.00	33.33	38.10	44.66	58.20	95.45
		2017	0.00	12.92	18.18	21.72	25.00	29.32	32.36	37.50	45.00	55.31	100.00
		2018	0.00	11.94	16.67	20.95	23.08	26.56	31.82	37.50	42.86	50.00	100.00
	LQS(III)	2015	0.00	17.50	22.22	26.32	30.11	35.55	40.82	46.44	55.25	65.36	100.00
		2016	0.00	12.66	19.41	23.53	28.57	32.53	36.84	40.76	46.15	57.87	100.00
		2017	0.00	13.33	18.35	22.72	26.22	29.63	33.76	39.54	44.99	54.83	100.00
		2018	0.00	7.12	16.67	21.44	25.00	27.27	32.26	37.90	42.22	50.00	100.00
0.3	LQS(I)	2015	0.00	9.72	14.57	18.75	22.55	26.14	30.62	36.02	43.15	54.55	100.00
		2016	0.00	6.73	13.33	17.90	20.34	23.62	27.41	32.48	40.00	49.29	100.00
		2017	0.00	4.09	12.71	16.67	20.00	24.06	28.89	33.33	38.24	45.00	90.00
		2018	0.00	5.18	13.79	16.83	19.35	22.45	25.37	28.57	36.36	47.34	100.00
	LQS(II)	2015	0.00	25.00	31.93	36.82	41.67	46.09	52.61	58.74	66.24	79.09	100.00
		2016	0.00	23.60	30.77	34.36	37.65	43.25	49.66	56.25	63.46	73.08	100.00
		2017	0.00	20.37	30.00	33.68	38.50	41.99	47.02	53.38	60.71	75.00	100.00
		2018	0.00	21.16	26.58	31.10	36.76	40.37	45.00	50.00	60.00	70.00	100.00
	LQS(III)	2015	0.00	28.78	36.19	40.87	46.46	53.91	60.71	67.64	74.08	83.41	100.00
		2016	0.00	23.78	31.62	37.19	42.94	48.78	54.41	60.04	67.06	78.67	100.00
		2017	0.00	25.10	30.72	36.21	40.00	44.31	48.94	57.14	64.16	74.76	100.00
		2018	0.00	21.63	28.57	33.33	38.30	42.41	47.69	54.42	60.38	71.43	100.00

Note that as defined above LQS is a percent indicator; the "%" symbol is omitted in the table cells

the $\delta = 0.1$ benchmark scenario is offered in the form of distinct yearly tree maps which are grouped in a *pdf* file available as Electronic Supplementary Material 2. As a sample illustration of the concrete underpinnings of our investigation Table 4 lists a selection of 6 JCR subject categories and presents their LQS indicators for each year, together with the other differentiation metrics..

Table 4 Sample of six JCR subject categories and their JIF quartiles differentiation metrics in the benchmark scenario of δ = 0.1 across four JCR editions

Category	JCR year	Absolute differences between JIF limiting quartile values				Poorly differentiated journals within each quartile pair and total unique poorly differentiated journals in the category					Indicators of local quartile similarity			% Share of non-meaningful pairwise differences (NPD)*
		Q1–Q2	Q2–Q3	Q3–Q4	Mean quartile difference	F(I) journals	F(II) journals	F(III) journals	Sum of unique F journals	% Share of unique F journals	LQS(I)	LQS(II)	LQS(III)	
Economics (SSCI)	2015	0.001	0.002	0.003	0.002	19	46	41	106	30.72	11.05	26.74	23.70	8.82
	2016	0.009	0.001	0.002	0.004	15	31	59	105	30.26	8.67	17.82	33.91	8.47
	2017	0.001	0.024	0.004	0.010	14	33	37	84	23.80	7.95	18.75	20.90	7.14
	2018	0.014	0.006	0.006	0.009	20	31	42	93	25.62	11.05	17.03	23.08	6.55
Geology (SCIE)	2015	0.001	0.074	0.005	0.027	4	2	8	14	29.79	17.39	8.33	33.33	7.77
	2016	0.247	0.097	0.037	0.127	0	2	7	9	19.15	0.00	8.33	29.17	8.60
	2017	0.343	0.021	0.022	0.129	0	5	7	12	25.53	0.00	20.83	29.17	6.75
	2018	0.070	0.056	0.008	0.045	2	4	9	15	31.91	8.70	16.67	37.50	8.05
History (SSCI)	2015	0.040	0.008	0.004	0.017	14	31	31	64	73.56	32.56	68.89	70.45	28.44
	2016	0.007	0.005	0.007	0.006	12	34	37	66	75.86	27.91	77.27	84.09	26.73
	2017	0.005	0.005	0.003	0.004	17	30	31	65	73.03	38.64	68.18	68.89	23.39
	2018	0.002	0.015	0.008	0.008	20	20	24	64	67.37	41.67	41.67	51.06	19.69
Information science & library science (SSCI)	2015	0.101	0.021	0.008	0.043	0	5	11	16	18.60	0.00	11.63	25.58	7.44
	2016	0.038	0.018	0.012	0.023	4	2	10	16	18.82	9.52	4.65	23.26	6.08
	2017	0.043	0.020	0.026	0.030	3	3	10	16	18.18	6.82	6.67	22.73	5.67
	2018	0.146	0.005	0.040	0.064	0	9	2	11	12.36	0.00	20.45	4.44	4.78
Logic (SCIE)	2015	0.043	0.031	0.003	0.026	7	10	7	14	63.64	63.64	90.91	63.64	32.03
	2016	0.096	0.058	0.020	0.058	3	5	3	11	52.38	30.00	55.56	27.27	17.62
	2017	0.053	0.009	0.001	0.021	4	7	8	14	70.00	40.00	70.00	80.00	24.74
	2018	0.099	0.004	0.019	0.041	2	9	5	13	65.00	20.00	90.00	50.00	25.79

Table 4 (continued)

Category	JCR year	Absolute differences between JIF limiting quartile values				Poorly differentiated journals within each quartile pair and total unique poorly differentiated journals in the category					Indicators of local quartile similarity			% Share of non-meaningful pairwise differences (NPD)*
		Q1–Q2	Q2–Q3	Q3–Q4	Mean quartile difference	F(I) journals	F(II) journals	F(III) journals	Sum of unique F journals	% Share of unique F journals	LQS(I)	LQS(II)	LQS(III)	
Oncology (SCIE)	2015	0.017	0.006	0.006	0.010	6	11	10	27	12.68	5.66	10.38	9.35	3.15
	2016	0.014	0.007	0.007	0.009	4	9	8	21	9.68	3.70	8.33	7.34	2.95
	2017	0.029	0.009	0.025	0.021	6	11	6	23	10.31	5.41	9.82	5.36	3.09
	2018	0.014	0.026	0.003	0.014	7	8	11	26	11.30	6.09	6.96	9.57	3.29

*See main text for details regarding this indicator

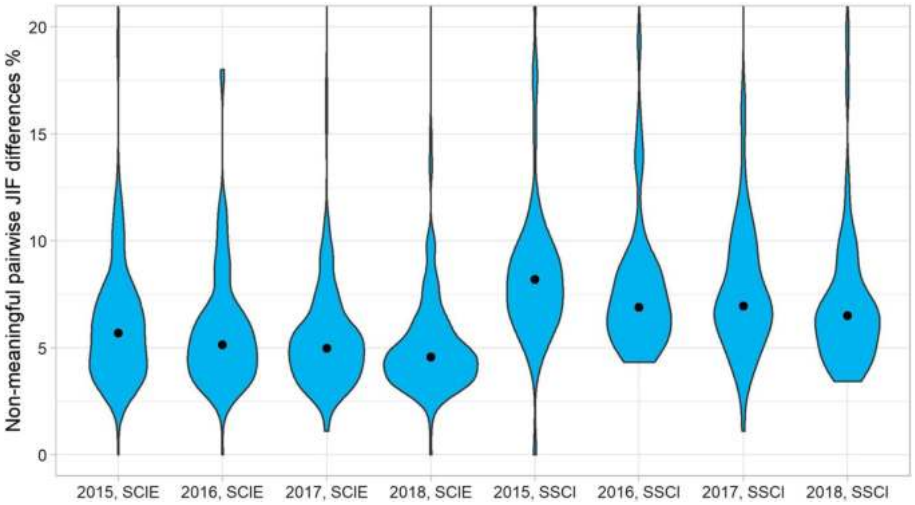

Fig. 6 Distribution of the SCIE and SSCI JCR categories by the share of non-meaningful JIF pairwise differences within all possible pairwise differences (NPD) between 2015 and 2018 with $\delta = 0.1$. The y axis is truncated at 20%; points mark median values

Additional exploration of quartile boundary differentiation and all pairwise JIF differences

In addition to our initial research questions we find it relevant to systematically describe the multiple relations between the quartile difference metrics discussed in the previous sections. We also find it highly relevant to underscore the fact that poor quartile boundary differentiation is in fact only a particular instance of a much more general problem of JIF dispersion within a subject category. We address this general problem in the current section and explore its relation to the share of poorly differentiated journals at the quartile boundaries within each category. We believe it reasonable to hypothesize that the JCR categories will show distinct quartile boundary homogeneity primarily in accordance with the overall dispersion of the complete set of underlying JIF values that originally determine the journal quartile assignment. Specifically, we expect fields with a higher overall share of non-meaningful JIF differences (i.e. a weak dispersion of JIF values) to also have a higher share of poorly differentiated journals at the quartile boundaries. Fields with a lower overall share of non-meaningful JIF differences (i.e. a higher JIF dispersion) should have a lower share of poorly differentiated quartile boundary journals.

To test our hypothesis we quantify JIF dispersion with an exhaustive approach that considers all possible pairwise differences between the JIF values in each JCR category. Given n journals in a category there are exactly $(n^2–n)/2$ possible pairwise differences. We assess if each of these differences is meaningful (under the three δ variations) and determine the characteristic percent share of *non-meaningful pairwise differences* (NPD) for

Table 5 Spearman rank-order correlations between differentiation metrics (2018 data, $\delta = 0.1$)

	Q1–Q2	Q2–Q3	Q3–Q4	F(I)	F(II)	F(III)	All F journals %	LQS(I)	LQS(II)	LQS(III)	NPD
Journals	–.55***	–.45***	–.57***	.63***	.72***	.74***	.01	.13*	–.03	.10	–.09
Q1–Q2		.29***	.39***	–.87***	–.46***	–.55***	–.40***	–.73***	–.08	–.26***	–.25***
Q2–Q3			.25***	–.40***	–.78***	–.40***	–.43***	–.18**	–.66***	–.11*	–.21***
Q3–Q4				–.42***	–.48***	–.80***	–.37***	–.19**	–.05	–.61***	–.13*
F(I)					.62***	.60***	.46***	.77***	.18**	.24***	.29***
F(II)						.65***	.46***	.26***	.59***	.20***	.28***
F(III)							.45***	.26***	.11*	.66***	.25***
All F journals %								.61***	.70***	.70***	.72***
LQS(I)									.24***	.30***	.45***
LQS(II)										.19**	.53***
LQS(III)											.46***

All p-values (*** for $p < .001$, ** for $p < .01$ and * for $p < .05$) correspond to directional hypotheses evaluated with single-tailed tests

each subject category.[7] We expect this share to be positively correlated with the percent share of poorly differentiated journals discussed in the previous section. Since we believe this indicator is of sufficient interest on its own we plot the corresponding distribution of JCR categories by their NPD values in Fig. 6.

Table 5 lists the Spearman correlation coefficients between the variables which informed our research questions, including the newly introduced NPD variable, in the context of the 2018 data under the benchmark $\delta = 0.1$ scenario. Very similar results are obtained for each of the other three JCR years and we use 2018 merely for exposition. Since the individual values of our variables are not normally distributed we opted for Spearman, rather than Pearson coefficients since the latter assume bivariate normality. Based on the results from the table absolute journal counts are negatively correlated with the size of the absolute differences between quartile boundaries (see also Fig. 2) and positively correlated with each F set, especially with F(II) and F(III) where coefficients higher than 0.7 are found. When looking to size-adjusted metrics however we find no relation between journal counts and the percent share of poorly differentiated journals or between journals and any of the three LQS indicators.[8]

Distinctive transversal patterns characterizing effect sizes may be detected within the correlation matrix: the quartile absolute differences Q1–Q2 are negatively and highly correlated with F(I) and with LQS(I) values. Q2–Q3 differences are highly and negatively correlated with F(II) and with LQS(II) values. Q3–Q4 differences are highly and negatively correlated with F(III) and with LQS(III) values. In addition, a 0.77 correlation describes the relation between the size-dependent F(I) and the size-independent LQS(I), a 0.59 correlation characterizes the relation between F(II) and LQS(II) and a 0.66 correlation characterizes the relation between F(III) and LQS(III). Aside from these transversal patterns, within variable group associations are also significant but weak in the case of the 3 absolute difference variables and also within the group of LQS indicators where coefficients lower than 0.3 can usually be observed. We find this important as it underscores the fact that poor quartile differentiation at one specific boundary is not necessarily related to the absence or presence of this phenomenon at the other two boundaries. Note for example some distinct LQS profiles based on Table 4: *geology* in 2017: 0, 21, 29; *history* in 2016: 28, 77, 84; *information science and library science* in 2018: 0, 20, 4.

Focusing on the percent share of (unique) poorly differentiated journals we observe that this is negatively, but only moderately correlated with each of the 3 absolute difference variables (coefficients of about –0.4) and positively but also only moderately correlated with the F set variables (coefficients of about 0.45). We find stronger associations between the percent share of poorly differentiated journals and each of the three individual LQS indicators, especially LQS(II) and LQS(III) where coefficients of 0.7 can be observed. Finally, considering the correlation with NPD we note that it offers fairly good support for the hypothesis that *poor differentiation at the quartile class boundaries is associated with poor overall JIF differentiation*: for the 2018 data significant although moderate correlations (coefficients between 0.45 and 0.53) exist between NPD and each of the three LQS indicators. However, between NPD and the percent share of poorly differentiated journals a

[7] For the fictitious category from Table 1 there are $(26^2 - 26) / 2 = 325$ JIF differences. Only 6.77% of these are not meaningful with $\delta = 0.1$. Some of these non-meaningful differences are precisely those at the quartile boundaries (Q1, Q2) and (Q3, Q4) which lead to the 23% poorly differentiated journals in the category.

[8] This can also be seen in the LQS profiles of the subject categories in Electronic Supplementary Material 2.

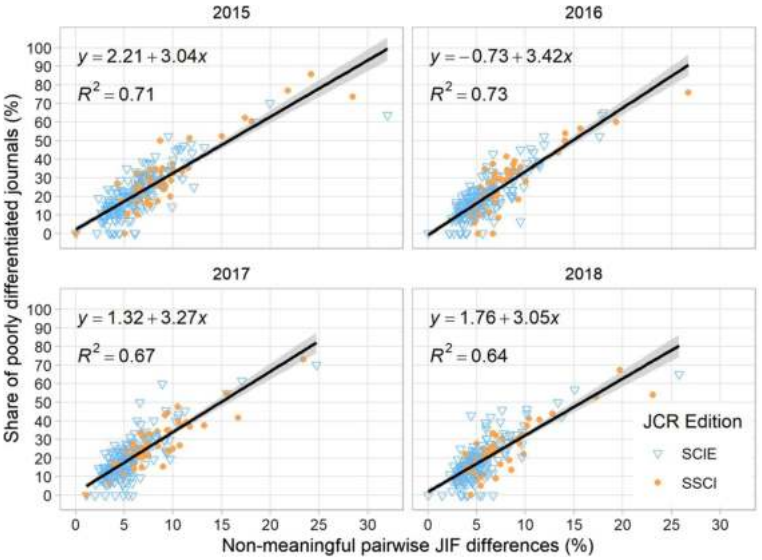

Fig. 7 Linear models of the share of poorly differentiated journals on the share of non-meaningful pairwise differences between JIFs (δ = 0.1). For all R^2 and slope estimates the p values are < .001; the intercept values are not significant

0.72 coefficient is obtained. For the other JCR years this correlation coefficient takes values between 0.77 and 0.8. Figure 7 presents a depiction of the relation between NPD and the percent share of poorly differentiated journals in the JCR subject categories for each year under the $\delta = 0.1$ benchmark. The figure illustrates that most of the variance in the share of poorly differentiated journals at the quartile boundaries seems to be explained by the overall share of non-meaningful pairwise differences.

Quartile temporal stability and poorly differentiated journals

We conclude our investigation with an analysis inspired by a reviewer who expressed interest in the problem of journal quartile stability. Specifically, we explore the stability of JIF quartiles throughout the four year period of our analysis and its connection to poor boundary differentiation. To this end, within each of the 234 subject categories, we determine the core journals, i.e. those that were consistently indexed in a category in each of the four consecutive years (matched across time by exact name). We then determine (1) the proportion of the core journals that experienced at least one quartile shift in the 4 consecutive years. Finally, for these latter journals we determine (2) the proportion of those that were also poorly differentiated at least once in the same period, under each of our 3 δ variations. We present in Fig. 8 the mapping of the JCR categories based on the 2 proportions just mentioned. Restricting our attention to the benchmark of $\delta = 0.1$ one last time we comment the first vertical pair of subplots from the figure. On the whole, there is a clear difference between SSCI and SCIE categories. The former typically have higher rates of journals with quartile shifts that also tend to suffer from poor boundary differentiation (note the concentration of points in the right

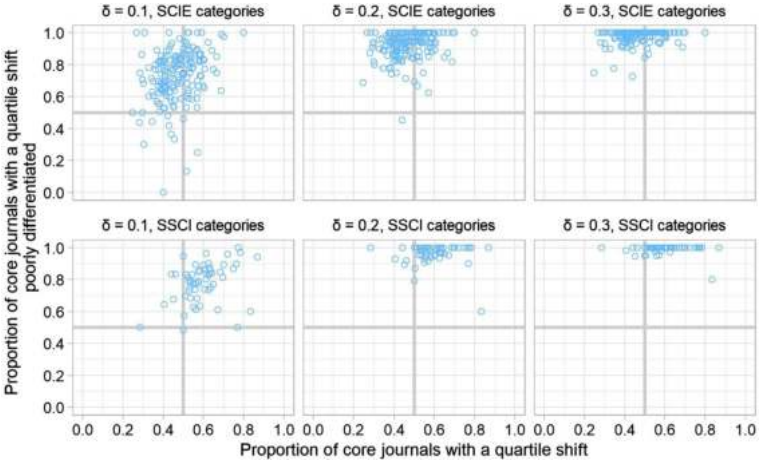

Fig. 8 JCR categories viewed with respect to the proportion of core journals that experienced a quartile shift at least once in the 2015–2018 period (x-axes) and the proportion of those journals that in addition were also poorly differentiated at least once in the same period (y-axes). Each vertical pair of subplots shows the SCIE/SSCI categories for a specific δ value. Note that across the SCIE and SSCI categories the δ variations only affect the y-component

upper region). The SCIE categories are less affected by quartile shifts but among the journals that do experience changes most also suffer from poor differentiation.

If we concentrate on the *x* component of our mappings we may note there is considerable temporal volatility for most JCR categories. Specifically, for all but seven of the 177 SCIE fields at least 30% of the core journals have experienced a quartile shift at least once in the 4 year window. For all but 5 of the 57 SSCI fields at least 50% of the core journals have experienced a quartile change at least once. Quartile volatility is thus visibly higher for social sciences compared to SCIE categories. For SSCI categories the median proportion of journals with at least one quartile shift is 0.58 while for SCIE categories it is 0.46. Among SCIE categories *logic* (with 0.80), *history & philosophy of science* (0.70) and *mathematics* (0.69) have the highest rates of quartile volatility, while *organic chemistry* (0.28), *legal medicine* (0.27) and *condensed matter physics* (0.25) have the lowest rates. For SSCI categories *cultural studies* (with 0.87), *green & sustainable science & technology* (0.83) and *history* (0.78) have the highest rates and *developmental psychology* (0.44), *psychiatry* (0.41) and *biological psychology* (0.29) have the lowest. Moving to the *y* component of the plots we can observe that for most categories, whether SCIE or SSCI, at least 60% of the core journals that experienced a quartile shift in the four years of our analysis also belonged to the group of poorly differentiated journals at least once in the same period. There is thus a substantial overlap between quartile shifts and poor differentiation. The overlap becomes striking when considering the sensitivity of the results to increased δ values and is documented in the second and third vertical pairs of subplots from Fig. 8. For most SCIE categories these show that under higher δ values at least 80% of the core journals with a quartile shift would be deemed poorly differentiated. For SSCI categories the same share exceeds 90%.

Discussion and conclusions

We have observed that the problem of poor quartile boundary differentiation seems to be decreasing discreetly with time. The results documented in Table 3 and Figs. 4, 5, 6 provide *sui generis* evidence in support of the phenomenon of impact factor inflation (Althouse et al. 2009; Chorus and Waltman 2016; Larivière and Sugimoto 2019) which essentially consists in JIFs gradually but consistently increasing over time. If one looks to the median LQS values (especially those obtained with δ of 0.2 and 0.3 which are more inclusive detectors of similarity) it is apparent that within the four year period studied in this paper they are continually decreasing. This may be attributable to the fact that JIF inflation is driving the quartile class boundary journals further apart. In addition, a closer inspection of the NPD indicator (with $\delta=0.1$) also highlights that between 2015 and 2018 78% of the JCR categories experienced a decrease in the share of within-field non-meaningful pairwise JIF differences.[9] It is not surprising that an increase in JIF values over time is associated with higher dispersion which, in turn, reduces similarity. However, further study of the evolution of JIFs across different subject categories over time is necessary in order to establish a clear relationship between time and a decrease in overall poor differentiation. We would expect this phenomenon to differ in intensity for the JCR categories and it might be interesting to investigate it for a much wider range of years.

In spite of JIF inflation our other empirical results show that the differences between the limiting values of JIF quartile classes remain in general very small and cannot credibly be described as minimally meaningful. As a rule, observing even the mildest requirement for a meaningful difference ($\delta=0.1$) leads to the conclusion that for the vast majority of JCR subject categories the boundary differences between classes are not meaningfully different. Poor differentiation around quartile boundaries is a legitimate caveat for research evaluation in which only the quartile classification scheme is employed as for about half of the JCR categories at least 20% of the journals are affected by poor differentiation and, for many, more than a third are affected. We stress that the singular use of JIF quartiles is pernicious in evaluations conducted at the micro level. For example the hiring and promotion of individual scholars should not be decided solely by consideration of JIF quartiles. Nonetheless, the quartiles may serve as legitimate heuristic devices in comparative studies at the macro level, for instance in comparing the scientific output of distinct countries.

Our results explain and reinforce a previous finding according to which in the yearly evolution of journal rankings "the most common changes are swaps between neighbouring quartiles" (Pajić 2015, p. 995). These swaps are a direct consequence of poor quartile boundary differentiation. The phenomena documented in our paper may also be related to a recent study which argues that comparing papers via JIFs—including papers from Q1 and Q2 journals—carries an inherent failure probability: "the result will be wrong when a JIF-sub-evaluated paper in a journal is compared with a JIF-over-evaluated paper in another journal" (Brito and Rodríguez-Navarro 2019, p. 315). The failure probabilities described in that study and supported with limited data on 6 SCIE JCR categories are closely linked to the poor differentiation of journals systematically documented in the present work. However, while we have shown in the preceding section that poor overall differentiation tends to imply poor differentiation at the quartile class boundaries, a variety of quartile similarity

[9] Although not our primary line of inquiry, we also looked at the evolution of mean JIF values in each JCR category between 2015 and 2018: all except 8 of the 234 categories that could be compared experienced an increase in the mean JIF, typically ranging between 12 and 43%.

profiles are observable for the different JCR categories. In the final analysis it is only by looking at each individual category that a rigorous, contextually relevant assessment can be made because poor differentiation affects some fields more than others. This is consistent with the fact that the citation density is heterogeneous across scientific disciplines (Lundberg 2007; van Raan 2019; Waltman 2016).

Earlier work regarding the reliability of JIF rankings in *research and experimental medicine* argued against their use and recommended "that a plausible range, such as a credible interval, be used alongside the estimated journal impact factor and its rank" (Greenwood 2007). In direct recognition of the vagueness of JIF quartiles García et al. (2012) used multidimensional fuzzy clustering as a way to address uncertainty in impact class assignment for *computer science—artificial intelligence*. For *economics* journals it was then shown that even for five-year JIFs there is considerable overlap of the confidence intervals that characterize the metrics of different journals (Stern 2013). The idea that confidence intervals are necessary for JIFs has also been advocated more recently as a way to counter the measurement errors inherent in bibliometric data (Bornmann 2017). JIF quartile classes contain *less* information than the original JIF values and the ranks derived from them. We believe it is evident that their independent use runs counter to the idea that uncertainty in bibliometric appraisal is a critical issue which should be confronted and mitigated rather than reinforced. Mitigating the JIF precision problem would require at least rounding to a single digit and an outspoken acceptance of the fact that this does lead to ties. There is no sound mathematical justification for vilifying these ties and for going to arbitrary lengths to artificially remove them, let alone for condoning their artificial enlargement in qualitative terms.

As noted in the introductory section we argue that the singular use of the JIF quartile system is an unfortunate second order ecological fallacy that compounds, rather than addresses the first order ecological fallacy of ascribing to a paper the impact merits of its publishing journal. This is so because in using JIF quartiles the perceived merits (or failings) of a quartile class are gratuitously transferred to the constituent journals and then further down to their individual papers, absent any evaluation of the papers themselves. Recently, in a study of 25 JCR SCIE categories (Miranda and Garcia-Carpintero 2019, p. 500) it has been shown that even "the relative quality significance of publishing papers in Q1 compared to other quartiles is largely dependent on the research area". An earlier comment (Liu et al. 2016) noted, somewhat paradoxically, that for SCIE journals Q1 papers are in fact the most common of papers simply because Q1 journals publish more articles than the journals in the other quartiles. Adding to these findings we believe that poor JIF quartile boundary differentiation should dissuade science administrators and policymakers from continuing to rely on the quartile system as the sole arbiter of scientific merit. This practice is unwarranted as a means of evaluating individual scholars, especially those working in social sciences.

In addition to the well-documented skewness of citation counts (Albarrán et al. 2011; Ruiz-Castillo and Costas 2018) one increasingly important aspect relevant for the evaluation of individual papers is the fact that there is a diminishing relation between JIFs and the citation rate of individual articles published in a journal as digital development changes the circulation of scholarly literature (Lozano et al. 2012). This is an additional reason to abandon the top-down approach of conflating a JIF with an individual research paper. As a better alternative it has been argued that the analysis of a paper's impact should preferably be conducted by studying it within the citation distribution of all articles published in a journal (Larivière et al. 2016). On the other hand, continued efforts are being devoted to creating more sophisticated JIF alternatives (e.g.: Leydesdorff et al. 2019) as well as to

creating new generation metrics that reflect more of a journal's functions than are captured by citation indicators such as JIFs (Wouters et al. 2019). Against this background it seems especially curious that an aggregated and simplifying indicator—the quartile classes built on the foundations of a problematic JIF—is still adopted in many contexts as a ready-made solution for the evaluation of scholars, regardless of their scientific field.

We conclude with a remark concerning the appealing visual rendition of boxplots which, as we mentioned, seem to have inspired the JIF quartiles. Despite their value for data analysis boxplots have been recognized as having limitations inherent not only in their design, but especially in their practical use. As early as 1978 it was known that "frequently, misinterpretation results because the viewer, particularly the nonstatistician, attempts to gain more information from the display than it contains" (McGill et al. 1978, p. 13). We would argue that this type of misinterpretation that follows from adding meaning where it is in fact not warranted is necessarily ingrained in the use of JIF quartiles as a singular evaluation device. The incontrovertible fact is that the information offered by the quartiles alone is simply insufficient and inadequate to inform a fair judgement of individual researchers.

Acknowledgements The authors express their gratitude to the anonymous reviewers whose comments helped to improve significant aspects of the initial manuscript. This paper was financially supported by the Human Capital Operational Program 2014-2020, co-financed by the European Social Fund, under the project POCU/380/6/13/124708 no. 37141/23.05.2019 with the title "Researcher-Entrepreneur on Labour Market in the Fields of Intelligent Specialization (CERT-ANTREP)", coordinated by the National University of Political Studies and Public Administration.

References

Adigozalova, N. A. (2019). Quartile weighted impact factor. *COLLNET Journal of Scientometrics and Information Management, 13*(2), 365–386. https://doi.org/10.1080/09737766.2020.1716646.

Aksnes, D. W. (2003). A macro study of self-citation. *Scientometrics, 56*(2), 235–246. https://doi.org/10.1023/A:1021919228368.

Albarrán, P., Crespo, J. A., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 subfields and a number of aggregates. *Scientometrics, 88*(2), 385–397. https://doi.org/10.1007/s11192-011-0407-9.

Althouse, B. M., West, J. D., Bergstrom, C. T., & Bergstrom, T. (2009). Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology, 60*(1), 27–34. https://doi.org/10.1002/asi.20936.

Archambault, É., & Larivière, V. (2009). History of the journal impact factor: Contingencies and consequences. *Scientometrics, 79*(3), 635–649. https://doi.org/10.1007/s11192-007-2036-x.

Bornmann, L. (2017). Confidence intervals for journal impact factors. *Scientometrics, 111*(3), 1869–1871. https://doi.org/10.1007/s11192-017-2365-3.

Bornmann, L., & Marx, W. (2014). How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations. *Scientometrics, 98*(1), 487–509. https://doi.org/10.1007/s11192-013-1161-y.

Brito, R., & Rodríguez-Navarro, A. (2018). Research assessment by percentile-based double rank analysis. *Journal of Informetrics, 12*(1), 315–329. https://doi.org/10.1016/j.joi.2018.01.011.

Brito, R., & Rodríguez-Navarro, A. (2019). Evaluating research and researchers by the journal impact factor: Is it better than coin flipping? *Journal of Informetrics, 13*(1), 314–324. https://doi.org/10.1016/j.joi.2019.01.009.

Campanario, J. M. (2014). The effect of citations on the significance of decimal places in the computation of journal impact factors. *Scientometrics, 99*(2), 289–298. https://doi.org/10.1007/s11192-013-1206-2.

Chorus, C., & Waltman, L. (2016). A large-scale analysis of impact factor biased journal self-citations. *PLoS ONE, 11*(8), e0161021. https://doi.org/10.1371/journal.pone.0161021.

Collier, K. (2019). Announcing the 2019 Journal Citation Reports. https://clarivate.com/webofsciencegroup/article/announcing-the-2019-journal-citation-reports/

Clarivate Analytics. (2018). *InCites Indicators Handbook*. http://help.incites.clarivate.com/inCites2Live/8980-TRS/version/default/part/AttachmentData/data/InCites-Indicators-Handbook - June 2018.pdf

Curry, S. (2018). Let's move beyond the rhetoric: It's time to change how we judge research. *Nature, 554*(7691), 147–147. https://doi.org/10.1038/d41586-018-01642-w.

Fernández-Ríos, L., & Rodríguez-Díaz, J. (2014). The "impact factor style of thinking": A new theoretical framework. *International Journal of Clinical and Health Psychology, 14*(2), 154–160. https://doi.org/10.1016/S1697-2600(14)70049-3.

García, J. A., Rodriguez-Sánchez, R., Fdez-Valdivia, J., & Martinez-Baena, J. (2012). On first quartile journals which are not of highest impact. *Scientometrics, 90*(3), 925–943. https://doi.org/10.1007/s11192-011-0534-3.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science, 178*(4060), 471–479. https://doi.org/10.1126/science.178.4060.471.

Garfield, E. (1990). How ISI selects journals for coverage: Quantitative and qualitative considerations. *Current Contents, 13*(22), 185–193.

Garfield, E. (2006). The History and meaning of the journal impact factor. *Journal of the American Medical Association, 295*(1), 90–93. https://doi.org/10.1001/jama.295.1.90.

Greenwood, D. C. (2007). Reliability of journal impact factor rankings. *BMC Medical Research Methodology, 7*(1), 48. https://doi.org/10.1186/1471-2288-7-48.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). The leiden manifesto for research metrics. *Nature, 520*(7548), 9–11. https://doi.org/10.1038/520429a.

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *American Statistician, 52*(2), 181–184. https://doi.org/10.1080/00031305.1998.10480559.

Larivière, V., Kiermer, V., MacCallum, C. J., McNutt, M., Patterson, M., Pulverer, B., et al. (2016). A simple proposal for the publication of journal citation distributions. *BioRxiv*. https://doi.org/10.1101/062109.

Larivière, V., & Sugimoto, C. R. (2019). The journal impact factor: A brief history, critique, and discussion of adverse effects. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 3–24). Cham: Springer.

Leydesdorff, L., & Bornmann, L. (2016). The operationalization of "fields" as WoS subject categories (WCs) in evaluative bibliometrics: The cases of "library and information science" and "science & technology studies." *Journal of the Association for Information Science and Technology, 67*(3), 707–714. https://doi.org/10.1002/asi.23408.

Leydesdorff, L., Bornmann, L., & Adams, J. (2019). The integrated impact indicator revisited (I3*): A non-parametric alternative to the journal impact factor. *Scientometrics, 119*(3), 1669–1694. https://doi.org/10.1007/s11192-019-03099-8.

Leydesdorff, L., Wouters, P., & Bornmann, L. (2016). Professional and citizen bibliometrics: Complementarities and ambivalences in the development and use. *Scientometrics, 109*(3), 2129–2150. https://doi.org/10.1007/s11192-016-2150-8.

Liu, W. (2020). China's SCI-indexed publications: Facts, feelings, and future directions. *ECNU Review of Education, 3*(3), 562–569. https://doi.org/10.1177/2096531120933902.

Liu, W., Hu, G., & Gu, M. (2016). The probability of publishing in first-quartile journals. *Scientometrics, 106*(3), 1273–1276. https://doi.org/10.1007/s11192-015-1821-1.

Lozano, G. A., Larivière, V., & Gingras, Y. (2012). The weakening relationship between the impact factor and papers' citations in the digital age. *Journal of the American Society for Information Science and Technology, 63*(11), 2140–2145. https://doi.org/10.1002/asi.22731.

Lundberg, J. (2007). Lifting the crown—citation z-score. *Journal of Informetrics, 1*(2), 145–154. https://doi.org/10.1016/j.joi.2006.09.007.

Madhan, M., Gunasekaran, S., Rani, M. T., Arunachalam, S., & Abinandanan, T. A. (2020). Chemistry research in India in a global perspective- A scientometrics profile, (February), 1–39. arXiv preprint arXiv:2002.03093v2.

Magri, M.-H., & Solari, A. (1996). The SCI journal citation reports: A potential tool for studying journals? *Scientometrics, 35*(1), 93–117. https://doi.org/10.1007/BF02018235.

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *The American Statistician, 32*(1), 12–16. https://doi.org/10.1080/00031305.1978.10479236.

McVeigh, M. E., & Mann, S. J. (2009). The Journal impact factor denominator. *Journal of the American Medical Association, 302*(10), 1107–1109. https://doi.org/10.1001/jama.2009.1301.

Milojević, S. (2020). Practical method to reclassify web of science articles into unique subject categories and broad disciplines. *Quantitative Science Studies, 1*(1), 183–206. https://doi.org/10.1162/qss_a_00014.

Miranda, R., & Garcia-Carpintero, E. (2019). Comparison of the share of documents and citations from different quartile journals in 25 research areas. *Scientometrics, 121*(1), 479–501. https://doi.org/10.1007/s11192-019-03210-z.

Müller, R., & de Rijcke, S. (2017). Thinking with indicators. exploring the epistemic impacts of academic performance indicators in the life sciences. *Research Evaluation, 26*(3), 157–168. https://doi.org/10.1093/reseval/rvx023.

Pajić, D. (2015). On the stability of citation-based journal rankings. *Journal of Informetrics, 9*(4), 990–1006. https://doi.org/10.1016/j.joi.2015.08.005.

Pudovkin, A. I., & Garfield, E. (2012). Rank normalization of impact factors will resolve Vanclay's dilemma with TRIF. *Scientometrics, 92*(2), 409–412. https://doi.org/10.1007/s11192-012-0634-8.

Quan, W., Chen, B., & Shu, F. (2017). Publish or impoverish. *Aslib Journal of Information Management, 69*(5), 486–502. https://doi.org/10.1108/AJIM-01-2017-0014.

R Core Team. (2020). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. https://www.r-project.org/.

Rafols, I., & Robinson-Garcia, N. (2016). On the dominance of quantitative evaluation in peripherall countries: Auditing research with technologies of distance. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2818335.

Ruiz-Castillo, J., & Costas, R. (2018). Individual and field citation distributions in 29 broad scientific fields. *Journal of Informetrics, 12*(3), 868–892. https://doi.org/10.1016/j.joi.2018.07.002.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ, 314*(7079), 497–497. https://doi.org/10.1136/bmj.314.7079.497.

Shu, F., Quan, W., Chen, B., Qiu, J., Sugimoto, C. R., & Larivière, V. (2020). The role of web of science publications in China's tenure system. *Scientometrics, 122*(3), 1683–1695. https://doi.org/10.1007/s11192-019-03339-x.

Stern, D. I. (2013). Uncertainty measures for economics journal impact factors. *Journal of Economic Literature, 51*(1), 173–189. https://doi.org/10.1257/jel.51.1.173.

Thelwall, M. (2016). Are the discretised lognormal and hooked power law distributions plausible for citation data? *Journal of Informetrics, 10*(2), 454–470. https://doi.org/10.1016/j.joi.2016.03.001.

van Raan, A. (2019). Measuring science: Basic principles and application of advanced bibliometrics. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 237–280). Cham: Springer.

Vanclay, J. K. (2009). Bias in the journal impact factor. *Scientometrics, 78*(1), 3–12. https://doi.org/10.1007/s11192-008-1778-4.

Vanclay, J. K. (2012). Impact factor: Outdated artefact or stepping-stone to journal certification? *Scientometrics, 92*(2), 211–238. https://doi.org/10.1007/s11192-011-0561-0.

Vîiu, G.-A. (2018). The lognormal distribution explains the remarkable pattern documented by characteristic scores and scales in scientometrics. *Journal of Informetrics, 12,* 401–415. https://doi.org/10.1016/j.joi.2018.02.002.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365–391. https://doi.org/10.1016/j.joi.2016.02.007.

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of web of science and scopus. *Journal of Informetrics, 10*(2), 347–364. https://doi.org/10.1016/j.joi.2016.02.003.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer.

Wilhite, A. W., & Fong, E. A. (2012). Coercive citation in academic publishing. *Science, 335*(6068), 542–543. https://doi.org/10.1126/science.1212540.

Wouters, P., Sugimoto, C. R., Larivière, V., McVeigh, M. E., Pulverer, B., de Rijcke, S., & Waltman, L. (2019). Rethinking impact factors: Better ways to judge a journal. *Nature, 569*(7758), 621–623. https://doi.org/10.1038/d41586-019-01643-3.

Zhu, J. (2020). Evaluation of scientific and technological research in China's colleges: A review of policy reforms, 2000–2020. *ECNU Review of Education, 3*(3), 556–561. https://doi.org/10.1177/2096531120938383.

Zijlstra, H., & McCullough, R. (2016). CiteScore: a new metric to help you track journal performance and make decisions. https://www.elsevier.com/editors-update/story/journal-metrics/citescore-a-new-metric-to-help-you-choose-the-right-journal