Modelling Continuous Percentile Rank Scores and Integrated Impact Indicators (I3) / Une modélisation des notations continues de classement par pourcentage et des indicateurs intégrés d'impact (I3)

Fred Y. Ye, Ronald Rousseau

Canadian Journal of Information and Library Science, Volume 37, Number 3, September / septembre 2013, pp. 201-206 (Article)

Published by University of Toronto Press
DOI: https://doi.org/10.1353/ils.2013.0016

Modelling Continuous Percentile Rank Scores and Integrated Impact Indicators (I3)

Une modélisation des notations continues de classement par pourcentage et des indicateurs intégrés d'impact (I3)

Fred Y. Ye
School of Information Management, Nanjing University, China
yye@nju.edu.cn

Ronald Rousseau
VIVES (Association KU Leuven), Faculty of Engineering Technology,
Oostende, Belgium
Universiteit Antwerpen (UA), IBW, Belgium
KU Leuven, Belgium
ronald.rousseau@uantwerpen.be

Abstract: Based on the well-known discrete definitions we introduce a continuous framework for percentile rank scores and integrated impact indicators (I3). This is done by taking the integral of a scoring function multiplied by a distribution function. Examples are provided by considering several distribution functions and two scoring functions, where the distribution function can take any form and the scoring function is non-decreasing.

Keywords: percentile rank scores; integrated impact indicator; I3; percentile; continuous modelling

Résumé : Sur la base des définitions distinctes bien connues, nous introduisons un cadre continu pour les notations de classement par pourcentage et les indicateurs intégrés d'impact (I3). Ceci est fait en pregnant l'intégrale d'une fonction de notation multipliée par une fonction de distribution. Des exemples sont fournis en tenant compte de plusieurs fonctions de distribution et de deux fonctions de notation, où la fonction de distribution peut prendre n'importe quelle forme et la fonction de notation est non décroissante.

Mots-clés : notations de classement par pourcentage; indicateur intégré d'impact; I3; centile; modélisation en continu

Introduction

Although researchers have realized that using arithmetic averages in scientometric investigations may lead to biased results (Leydesdorff et al. 2011), it has taken several years before an acceptable alternative was formulated. Slowly a

consensus has arisen, leading to the use of percentiles and percentile rank classes (Bornmann 2010; Bornmann et al. 2013; Leydesdorff and Bornmann 2011; Leydesdorff et al. 2011; Opthof and Leydesdorff 2010; Rousseau 2012). These notions are based on the concept of percentiles (or quantiles) for discrete data. As most informetric models can also be described within a continuous context (Egghe 2005), we propose a continuous analogue of the percentile approach and, as an illustration, calculate the resulting percentile rank scores and Integrated Impact Indicator (see further for definitions) for some basic functions.

Definitions

In this section, we use the framework as presented by Rousseau (2012). Consider a set A and a reference set S containing all elements in A, hence $A \subseteq S$. Moreover, we assume that a function X from S to the positive real numbers is given, leading to the image multiset $X(S)$. Note that we consider $X(S)$ as a multiset, as we consider the images $X(s)$, s in S, as separate entities (even if their values are the same). A standard situation is the case that A consists of a set of articles, set S consists of all articles in the journals in which set A is published (published in the same year), and a function X which maps an article to the number of citations it has received over a given period (and there may be several articles with the same number of citations).

Now a rule is given which subdivides set S into M disjoint classes, based on the values of the function X. If a document belongs to class m, then it receives a score x_m. Note that this score only depends on the class (and hence on S), but may not depend on set A (Rousseau and Ye 2012). Again a standard situation is the case that there are 100 percentile classes (or 10 decile classes). In the case of percentiles, articles belonging to the top 1% receive a score of 100; those belonging to the top 2% (and not to the top 1%) receive a score of 99, and so on. Besides classes of equal breadth, one may also use classes of unequal breadth such as the six US National Science Foundation categories (National Science Board 2010).

Definition 1. Percentile rank scores (Bornmann and Mutz 2011; Leydesdorff et al. 2011)

Let A be a set of N documents, assume there are M classes, and let $n_A(m)$ be the number of documents in A that belong to class m. Then the percentile rank score of A is defined as:

$$R(A) = \sum_{m=1}^{M} x_m \frac{n_A(m)}{N} \tag{1}$$

$R(A)$ can be seen as a weighted average of scores. Clearly, the value of $R(A)$ depends not only on A, but also on the reference set S, the M classes used, and their score. We note that this indicator allows a lot of flexibility, but hence also a lot of subjectivity, as one can adapt the reference set, the classes, and the scores.

Definition 2. The Integrated Impact Indicator (I3)
The I3 indicator (Leydesdorff and Bornmann 2011), where I3 stands for Integrated Impact Indicator, is defined in a similar way as the percentile rank score as given in Equation (1). The role of the reference set S is the same, but this time, no division by N is performed. Hence, using the notation introduced earlier, we have the following definition.

Definition 3. The I3 score of a set A *is defined as:*

$$\mathrm{I3}(A) = \sum_{m=1}^{M} x_m n_A(m) \tag{2}$$

Clearly, $\mathrm{I3}(A) = N \cdot R(A)$.

In the context of journal impact, I3 is preferred to R as "having an impact" implies publishing many articles *and* receiving many citations.

We generalize the step functions (in Equation (2)) to continuous functions $w(x)$ and $k(x)$ defined on an interval $[0,C]$, $C > 0$. This leads to:

$$R = \int_0^C w(x) \cdot k(x)\mathrm{d}x \tag{3}$$

The function $k(x)$ is a density function, and $w(x) > 0$ is a scoring function acting as a weight for the function $k(x)$. The origin of the interval $[0,C]$ corresponds to the worst results—and hence the lowest scores—while the end point C corresponds to the best results and hence the highest scores. Consequently, $w(x)$ is a non-decreasing (usually strictly increasing) function, while the density function $k(x) \geq 0$ can have any form. If $f(x)$ is a positive integrable function on $[0,C]$, then we denote by N the integral $\int_0^C f(x)\mathrm{d}x$ and $f(x)/N$ becomes a density function on $[0,C]$. For any $k(x) = f(x)/N$, R (as defined in Equation (3)) times N becomes the continuous analogue of I3, which we also denote as I3. Hence, in a continuous setting $\mathrm{I3} = \int_0^C w(x) \cdot f(x)\mathrm{d}x$, where $f(x)$ is a positive integrable function on $[0,C]$.

The reason that we refer to our approach as a continuous approach is that, besides the continuous density function $k(x)$, we also consider a continuous analogue of the discrete weight or scoring values x_m.

Continuous examples: A first scoring function

We first consider the simple case that $w(x)$ is a linearly increasing function on the interval $[0,C]$:

$$w(x) = ax + \mathrm{c} \tag{4}$$

where c is a constant and $a > 0$ (as w is an increasing function). Choosing a value zero at the begin point, $w(0) = 0$, leads to $c = 0$; hence

$$w(x) = ax; a > 0 \tag{5}$$

Now we discuss different basic forms for the function $k(x)$.

Case 1. k(x) *is a constant, corresponding to a uniform distribution*
If $f(x) = K$ (a constant) then $N = K \cdot \mathrm{C}$ and $k(x) = \frac{1}{C}$ is a density function on [1,C], then we obtain

$$R = \int_0^C a \cdot x \cdot \frac{1}{C} \mathrm{d}x = \frac{a \cdot C}{2} \tag{6}$$

and

$$\mathrm{I3} = R \cdot N = \frac{a \cdot C^2 \cdot K}{2} \tag{7}$$

Case 2. k(x) *is a linear function*
We consider the linear decreasing function $f(x) = m\,(C - x)$, with $m > 0$ and $f(C) = 0$. Note that we take a linearly decreasing function because it is assumed here that there are many poor cases and few better ones. Normalizing yields $N = \int_0^C m \cdot (C - x)\mathrm{d}x = \frac{m \cdot C^2}{2}$ and hence $k(x) = \frac{2 \cdot (C - x)}{C^2}$ is a density function on [0,C].

Based on Equation (3) we obtain:

$$R = \int_0^C \frac{2a \cdot x \cdot (C - x)}{C^2}\,\mathrm{d}x = \frac{a \cdot C}{3} \tag{8}$$

and hence:

$$\mathrm{I3} = R \cdot N = \frac{a \cdot m \cdot C^3}{6} \tag{9}$$

Case 3. The function k(x) *is an exponential function*
We consider the function $f(x) = be^{mx}; m \neq 0; b > 0$. Then $\mathrm{N} = \int_0^C b \cdot e^{m \cdot x}\mathrm{d}x = \frac{b}{m}(e^{m \cdot C} - 1)$, leading to the density function $k(x) = \frac{m \cdot e^{m \cdot x}}{e^{m \cdot C} - 1}$. Then:

$$R = \int_0^C a \cdot x \cdot \frac{m \cdot e^{m \cdot x}}{e^{m \cdot C} - 1}\mathrm{d}x = \frac{a \cdot ((C \cdot m - 1) \cdot e^{C \cdot m} + 1)}{m \cdot (e^{C \cdot m} - 1)} \tag{10}$$

and hence:

$$\mathrm{I3} = \frac{a \cdot b \cdot ((C \cdot m - 1) \cdot e^{C \cdot m} + 1)}{m^2} \tag{11}$$

Case 4. The function k(x) *is a decreasing power function*
We consider $f(x) = \frac{1}{x+m}, m > 0$. Then $N = \int_0^C \frac{1}{x+m}\,\mathrm{d}x = \ln\left(\frac{C+m}{m}\right)$, and hence $k(x) = f(x)/N$ is a density function. This leads to:

$$R = \frac{a}{N}\int_0^C \frac{x}{x+m}\mathrm{d}x = \frac{a \cdot C}{\ln\left(\frac{C+m}{m}\right)} - a \cdot m \tag{12}$$

and

$$\mathrm{I3} = a \cdot C - a \cdot m \cdot \ln\left(\frac{C+m}{m}\right) \tag{13}$$

Case 5. k(x) *is a triangular peak function*
If $f(x)$ is a triangular peak function with peak point at $(C/2, b\ C/2)$, this yields:

$$f(x) = \begin{cases} b \cdot x, & 0 \le x \le \frac{C}{2} \\ b \cdot (C - x), & \frac{C}{2} < x \le C \end{cases} \tag{14}$$

Normalizing $f(x)$ leads to $N = b^2 \cdot C/4$ and the corresponding density function $k(x) = f(x)/N$. The integral for R consists of two parts:

$$R = \frac{4}{b^2 \cdot C}\int_0^{C/2} a \cdot b \cdot x^2\mathrm{d}x + \frac{4}{b^2 \cdot C}\int_{C/2}^{C} a \cdot b \cdot x(C-x)\mathrm{d}x$$

$$= \frac{a \cdot C^2}{6b} + \frac{a \cdot C^2}{3b} = \frac{a \cdot C^2}{2b} \tag{15}$$

leading to:

$$\mathrm{I3} = \frac{a \cdot b \cdot C^3}{8} \tag{16}$$

Continuous examples: A second scoring function
Finally we consider a second scoring function $w(x)$ which increases faster than a linear function. We consider the following increasing power function $w(x) = a \cdot x^{\alpha}, a > 0, \alpha > 1$. For $f(x)$ we take the function $b \cdot x^{\beta}$, with $b > 0$. Then $N = \int_0^C b \cdot x^{\alpha}\mathrm{d}x = \frac{b}{\beta+1} C^{\beta+1}$. This leads to:

$$R = \frac{1}{N}\int_0^C a \cdot x^{\alpha} \cdot b \cdot x^{\beta}\mathrm{d}x = \frac{a \cdot (\beta+1) \cdot C^{\alpha}}{\alpha+\beta+1} \text{ and } \mathrm{I3} = \frac{a \cdot b \cdot C^{\alpha+\beta+1}}{\alpha+\beta+1} \tag{17}$$

Conclusion

We calculated the value of I3 in a continuous framework for different distribution functions $k(x)$ and for two scoring functions $w(x) = ax$ and $w(x) = a \cdot x^{\alpha}$. In this way we introduced a method by which R and the I3 indicator can be

used in a continuous modelling context. In all cases, the resulting values are functions of the parameters introduced by the functions $w(x)$ and $k(x)$. One reviewer correctly pointed out that this flexibility may lead to some additional difficulties if one wants to use a continuous approach for modelling real data. In such cases, $w(x)$, $k(x)$, and possibly C must be estimated. However, when working on this article we had an abstract theoretical framework in mind, namely one not based on real data. By proposing this continuous approach, we hope to stimulate further investigations within a continuous modelling approach to I3.

Acknowledgements

The authors acknowledge the National Natural Science Foundation of China (NSFC Grants No. 7101017006 and 71173187) for financial support. They thank their colleagues Wolfgang Glänzel, Leo Egghe, and Loet Leydesdorff for useful comments. Finally, anonymous reviewers and the editor are acknowledged for their useful suggestions to improve the manuscript.

References

Bornmann, Lutz. 2010. "Towards an Ideal Method of Measuring Research Performance: Some Comments to the Opthof and Leydesdorff (2010) paper." *Journal of Informetrics* 4 (3): 441–43. http://dx.doi.org/10.1016/j.joi.2010.04.004.

Bornmann, Lutz, Loet Leydesdorff, and Rüdiger Mutz. 2013. "The Use of Percentiles and Percentile Rank Classes in the Analysis of Bibliometric Data: Opportunities and Limits." *Journal of Informetrics* 7 (1): 158–65. http://dx.doi.org/10.1016/j.joi.2012.10.001.

Bornmann, Lutz, and Rüdiger Mutz. 2011. "Further Steps towards an Ideal Method of Measuring Citation Performance: The Avoidance of Citation (ratio) Averages in Field-Normalization." *Journal of Informetrics* 5 (1): 228–30. http://dx.doi.org/10.1016/j.joi.2010.10.009.

Egghe, Leo. 2005. *Power Laws in the Information Production Process: Lotkaianinformetrics*. Amsterdam: Elsevier.

Leydesdorff, Loet, and Lutz Bornmann. 2011. "Integrated Impact Indicators Compared with Impact Factors: An Alternative Research Design with Policy Implications." *Journal of the American Society for Information Science and Technology* 62 (11): 2133–46. http://dx.doi.org/10.1002/asi.21609.

Leydesdorff, Loet, Lutz Bornmann, Rüdiger Mutz, and Tobias Opthof. 2011. "Turning the Tables on Citation Analysis One More Time: Principles for Comparing Sets of Documents." *Journal of the American Society for Information Science and Technology* 62 (7): 1370–81. http://dx.doi.org/10.1002/asi.21534.

National Science Board. 2010. "*Science and Engineering Indicators*." Washington DC: National Science Foundation. http://www.nsf.gov/statistics/seind10/.

Opthof, Tobias, and Loet Leydesdorff. 2010. "Caveats for the Journal and Field Normalizations in the CWTS ("Leiden") Evaluations of Research Performance." *Journal of Informetrics* 4 (3): 423–30. http://dx.doi.org/10.1016/j.joi.2010.02.003.

Rousseau, Ronald. 2012. "Basic Properties of Both Percentile Rank Scores and the I3 Indicator." *Journal of the American Society for Information Science and Technology* 63 (2): 416–20. http://dx.doi.org/10.1002/asi.21684.

Rousseau, Ronald, and Fred Y. Ye. 2012. "A Formal Relation between the h-index of a Set of Articles and Their I3 score." *Journal of Informetrics* 6 (1): 34–35. http://dx.doi.org/10.1016/j.joi.2011.09.004.