The Problem of Percentile Rank Scores Used With Small Reference Sets

Dear Sir,

Instead of a relative mean citation rate, a percentile rank score (PRS) can be used in bibliometrics to generate a normalized citation impact for a paper. The use of a PRS is very advantageous, as no assumptions have to be made as to the distribution of citations in the reference set; that is, the scores are applicable also for the (usually) right-skewed distributions. To calculate the PRS for a single paper, a publication set (the reference set) must be compiled from a database (e.g., Thomson Reuters’ Web of Science) published with the same document type, in the same subject category (or journal), and in the same publication year.

To calculate the PRS, two steps are necessary: First, all publications in the set are ranked in increasing order $X_1 \leq X_2 \leq \ldots \leq X_n$ where $X_1$ ($X_n$) denotes the number of citations received by the least (most) cited publication. Second, each publication is assigned a PRS based on this distribution. If, for example, a single publication had 50 citations, and this citation count was equal to or greater than the citation counts of 90% of all publications, then the PRS of this publication would be at best 90. The publication would be in the 90th percentile (or higher). Because all publications in the reference set are attributed to a PRS, the PRS is also assigned to the publication in question. This PRS shows the publication’s normalized citation impact relative to the papers with the same document type and published in the same subject category (or journal) and year.

In response to the paper of Leydesdorff, Bornmann, Mutz, and Opthof (2011) several methods have been proposed to calculate the PRS, with the intention of finding a solution for the problem of small publication sets (Rousseau, 2012; Schreiber, 2012). For example, a reference set may contain a limited number of, say, ten papers. In that case, the highest possible PRS would be based on (9/10) or 90%. Rousseau (2012) proposes to define this highest possible rank as 100% by including the ranked paper in the ranking, and thus to consider (10/10) as the highest possible rank (see Leydesdorff, in press).

The calculation of PRS for small data sets is a well-known problem in statistics and several solutions have been proposed (Hyndman & Fan, 1996; Sheskin, 2007). The calculation of PRS using linear interpolation is integrated into statistical software packages (e.g., it is part of the EXAMINE procedure of SPSS, of the R-function QUANTILE, and of the CENTILE command of Stata). However, I would like to point out that reliable results in the calculation of PRS can only be reached if reference sets are used that are based on a larger publication set (journal sets or larger subject categories instead of single journals at best). Also, by using large sets the differences between the various methods for the proposed PRS calculation might be of little or no practical consequence. Small reference sets lead to unreliable performance estimations for single publications. Although the different proposals for dealing with small reference sets appear to be interesting solutions for this problem, the missing reliability produced by the low publication number remains in each case.

Thus, I recommend a use of PRS, which is oriented towards reference sets including sufficient publication numbers, instead of mathematical solutions for dealing with small paper numbers.

Yours sincerely,
Lutz Bornmann

References


Lutz Bornmann

Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstraße 8, 80539 Munich, Germany.
E-mail: bornmann@gv.mpg.de

Published online in Wiley Online Library (wileyonlinelibrary.com).