Journal of Informetrics 9 (2015) 642–657

Contents lists available at ScienceDirect

Journal of Informetrics

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

Predicting the long-term citation impact of recent publications☆

Clara Stegehuis[a,*], Nelly Litvak[b], Ludo Waltman[c]

[a] Eindhoven University of Technology, Department of Mathematics and Computer Science, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
[b] University of Twente, Department of Applied Mathematics, P.O. Box 217, 7500 AE Enschede, The Netherlands
[c] Centre for Science and Technology Studies, Leiden University, P.O. Box 905, 2300 AX Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 30 March 2015
Received in revised form 14 June 2015
Accepted 15 June 2015

Keywords:
Citation analysis
Citation impact
Impact factor
Prediction
Quantile estimation
Quantile regression

ABSTRACT

A fundamental problem in citation analysis is the prediction of the long-term citation impact of recent publications. We propose a model to predict a probability distribution for the future number of citations of a publication. Two predictors are used: the impact factor of the journal in which a publication has appeared and the number of citations a publication has received one year after its appearance. The proposed model is based on quantile regression. We employ the model to predict the future number of citations of a large set of publications in the field of physics. Our analysis shows that both predictors (i.e., impact factor and early citations) contribute to the accurate prediction of long-term citation impact. We also analytically study the behavior of the quantile regression coefficients for high quantiles of the distribution of citations. This is done by linking the quantile regression approach to a quantile estimation technique from extreme value theory. Our work provides insight into the influence of the impact factor and early citations on the long-term citation impact of a publication, and it takes a step toward a methodology that can be used to assess research institutions based on their most recently published work.

1. Introduction

Citation counts are a popular indicator of the impact of scientific publications. In the evaluation of research institutions, bibliometric indicators based on the citations received by the publications of an institution often play an important role. However, the use of citation-based indicators is problematic when the impact of recent publications needs to be determined. One or two years after their appearance, most publications have received only a few citations. After one year, there are many publications with just one or two citations or even with no citations at all. Some of these publications may receive a lot of citations in later years, while others may attract hardly any attention in the future. This makes it difficult to determine the impact of recent publications. Nevertheless, research institutions often want their performance to be assessed based on their most recent work (Bornmann, 2013). For example, a research institution may be interested in comparing the performance of different research groups based on publications from the four or five most recent years. Publications from earlier years may not be considered of interest anymore because many of them were produced by researchers who are no longer affiliated to

☆ The peer review process of this paper was handled by Vincent Larivière, Associate Editor of Journal of Informetrics.

* Corresponding author.
E-mail address: c.stegehuis@tue.nl (C. Stegehuis).

http://dx.doi.org/10.1016/j.joi.2015.06.005
1751-1577/

the institution or because they belong to old research lines that in the meantime have been closed down. Given the practical importance of taking into account very recent publications in research evaluations, our aim in this paper is to introduce a model for making predictions of the impact that recent publications will have in the long term.

Our model predicts the long-term citation impact of a publication based on two variables, namely the impact factor of the journal in which the publication has appeared and the number of early citations the publication has received. Early citations are defined as citations received in the year in which a publication appeared or in the year thereafter. The two predictors that we use are easily available, and contrary to for example the prediction approach proposed by Wang, Song, and Barabási (2013), they allow predictions to be made fairly soon after the appearance of a publication. Also, compared with other predictors that could be considered, such as the length of the reference list of a publication or the number of authors of a publication, the predictors that we use are relatively hard to manipulate. Earlier studies have shown that both impact factors and early citations are important predictors of future citations. In the next section, we will provide an overview of these earlier studies and we will discuss their relationship with our present work.

Earlier studies on citation impact prediction have often focused on providing a point estimate of the future number of citations of a publication. Given the high degree of uncertainty in citation impact predictions, we believe that it is more relevant to know the probability that a publication will receive a certain number of citations in the future. We therefore do not predict the average number of citations that a publication is expected to attract in the future, but instead we predict a probability distribution for the future number of citations of a publication. To predict this probability distribution conditional on a publication's impact factor and its early citations, we employ the technique of quantile regression introduced by Koenker and Bassett (1978).

We also study the relationship between our prediction model based on quantile regression and results from extreme value theory. To do so, we first use so-called Zenga plots, introduced recently by Cirillo (2013), to establish that the citation distributions obtained in our analysis have a Pareto tail. This result then enables us to provide analytical insight into the behavior of the quantile regression coefficients for high quantiles. More specifically, we are able to link the regression coefficients to an estimator for the tail quantiles of a Pareto distribution developed in the framework of extreme value theory (Dekkers, Einmahl, & De Haan, 1989).

We use citation data for a large set of publications in the field of physics to test our prediction approach. The data is taken from the Web of Science database.

The paper is organized as follows. First, Section 2 discusses how our research relates to earlier work reported in the literature. Next, Section 3 describes the data that were used in our analysis. Section 4 then introduces our model for predicting the long-term citation impact of publications, conditional on impact factors and early citations. Section 5 presents our empirical results. Sections 5.1 and 5.2 focus on the values obtained for the parameters of our model. Sections 5.3–5.6 address the fit of this model to the data and the predictive power of the model. Section 6 studies the relationship between our model and results from extreme value theory. Section 7 addresses the sensitivity of the parameters of our model to differences between fields of science, focusing on the fields of biology, chemistry and physics. Finally, Section 8 concludes the paper.

2. Relation with earlier work

There is an extensive literature on modeling or predicting the number of citations of a publication based on all kinds of variables. An early study in this literature is the work by Peters and Van Raan (1994), who investigate the determinants of the citation impact of chemical engineering publications. More recent work in this literature is reported by, among others, Bornmann, Leydesdorff, and Wang (2013), Didegah and Thelwall (2013a, 2013b), Fu and Aliferis (2010), Haslam et al. (2008), Walters (2006), Wang et al. (2012), Wang, Yu, and Yu (2011), Yu, Yu, Li, and Wang (2014), and Onodera and Yoshikane (2015). Various studies have also appeared in non-bibliometric journals (e.g. Haslam & Koval, 2010; Lokker, McKibbon, McKinlay, Wilczynski, & Haynes, 2008; Mingers & Xu, 2010). Recent overviews of the literature on modeling or predicting citation impact are provided by Didegah and Thelwall (2013a, 2013b) and Onodera and Yoshikane (2015). Examples of variables that have been found to predict citation impact include the impact factor of the journal in which a publication has appeared, the type of study (e.g., original research vs. literature review), the number of pages of a publication, the number of references of a publication, the number of authors, institutions, and countries in a publication's address list, and the past performance of these authors, institutions, and countries.

It is important to emphasize that the objective of our work is different from the studies mentioned above. Like the above-mentioned studies, our interest is in predicting citation impact. However, our more specific interest is in using citation impact predictions in the evaluation of researchers, research groups, research institutions, and so on. In this specific context, many of the variables that have been found to correlate with citation impact should not be used for making citation impact predictions. Some variables have the problem that they can be easily manipulated. For example, suppose researchers know that they will be evaluated based on the predicted citation impact of their publications, and suppose researchers also know that the citation impact of a publication will be predicted based on, for example, the number of pages or the number of references of the publication. In that case, in order to be evaluated more favorably, it might be tempting for researchers to try to artificially increase the number of pages or the number of references of their publications. Therefore, we consider only variables that cannot be altered easily by the authors of a publication. Other variables have the problem that they may lead to an undesirable self-reinforcing effect. For example, suppose researchers are evaluated based on the predicted