Journal of Informetrics 11 (2017) 713–729

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

Predicting the age of researchers using bibliometric data☆

Gabriela F. Nane[a,*], Vincent Larivière[b], Rodrigo Costas[c]

[a] Delft Institute of Applied Mathematics, Delft University of Technology, The Netherlands
[b] Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, Canada
[c] Center for Science and Technology (CWTS), Leiden University, The Netherlands

ARTICLE INFO

Article history:
Received 28 November 2016
Received in revised form 3 May 2017
Accepted 8 May 2017
Available online 15 June 2017

ABSTRACT

The age of researchers is a critical factor necessary to study the bibliometric characteristics of the scholars that produce new knowledge. In bibliometric studies, the age of scientific authors is generally missing; however, the year of the first publication is frequently considered as a proxy of the age of researchers. In this article, we investigate what are the most important bibibliometric factors that can be used to predict the age of researchers (birth and PhD age). Using a dataset of 3574 researchers from Québec for whom their Web of Science publications, year of birth and year of their PhD are known, our analysis falls under the linear regression setting and focuses on investigating the predictive power of various regression models rather than data fitting, considering also a breakdown by fields. The year of first publication proves to be the best linear predictor for the age of researchers. When using simple linear regression models, predicting birth and PhD years result in an error of about 3.7 years and 3.9 years, respectively. Including other bibliometric data marginally improves the predictive power of the regression models. A validation analysis for the field breakdown shows that the average length of the prediction intervals vary from 2.5 years for Basic Medical Sciences (for birth years) up to almost 10 years for Education (for PhD years). The average models perform significantly better than the models using individual observations. Nonetheless, the high variability of data and the uncertainty inherited by the models advice to caution when using linear regression models for predicting the age of researchers.

1. Introduction

Several sociodemographic factors have been shown to affect researchers' scholarly output and impact (Costas & Bordons, 2011; Gingras, Larivière, Macaluso, & Robitaille, 2008; Mauleón & Bordons, 2006). Among those, we can mention age (Costas & Bordons, 2011; Gingras et al., 2008; Levin & Stephan, 1989), gender (Larivière, Gingras, Cronin, & Sugimoto, 2013; Mauleón & Bordons, 2006), mobility and migration (Canibano, Otamendy, & Solis, 2011; Franzoni, Scellato, & Stephan, 2012; Moed & Halevi, 2014).

The development of large scale author-name disambiguation algorithms (Caron & Van Eck, 2014), as well as the increasing quantity of indexed papers' metadata (e.g. author names and surnames, affiliations, e-mail data, etc.) have expanded the possibilities to study such sociodemographic variables. For example, the analysis of the first author names of authors

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

* Corresponding author.
E-mail address: g.f.nane@tudelft.nl (G.F. Nane).

http://dx.doi.org/10.1016/j.joi.2017.05.002
1751-1577/

(Larivière et al., 2013) allowed for the macro analysis of gender disparities worldwide. The large-scale analysis of the relationship between author names, affiliations and countries has also opened the possibility of studying academic migrations at the world level (Moed, Aisati, & Plume, 2013), as well as the nationality (Costas & Noyons, 2013) or even the ethnic origin (Freeman, 2014) of scholars.

One of the central sociodemographic characteristics of scholars is their age (Costas & Bordons, 2011; Gingras et al., 2008; Levin & Stephan, 1989), as it has been shown to be a key predictor of research productivity (Bornmann & Leydesdorff, 2014; Falagas et al., 2008; Levin & Stephan, 1989). However, such variable is generally not included in bibliometric analyses, given its lack of availability. While several analyses have used the year of first publication as a proxy for their age, of a scholar (e.g. Radicchi & Castellano, 2013), there has not been any analysis on the actual relationship between this proxy and the real age of scholars. This paper is intended to fill this gap and shed some light on the underlying relationship between the 'bibliometric' age of scholars and their 'real' ages, defined as their biological age and time to PhD. In other words, we aim to assess how reliable is the estimation of the real ages of scholars based on models that exclusively rely on bibliometric indicators, such as the year of first publication, author order, co-authors, document types published, etc.

Firstly, we will investigate the correlations between all the variables considered in the analysis. Furthermore, several boxplots of the birth and PhD year will be presented and analysed in order to study the dispersion of the actual data. The next step in our analysis will focus on linear regression model fitting.[1] Therefore the birth (BIRTH hereafter) and PhD (PHD hereafter) years will be most frequently referred to as the 'dependent variables', while the bibliometric variables will be interchangeably referred to as the 'independent variables', covariates or predictors.

2. Methodology

For the study proposed it is absolutely necessary to have a dataset of scholars for whom the real ages of all the individuals considered are certainly known as well, as the publication years of their scientific publications, conforming the 'golden set' of the study. As golden set we have considered one of the (possibly) largest datasets of individual scholars for whom their actual individual characteristics are known (this dataset has been used in some previous studies, e.g. Gingras et al., 2008; Larivière et al., 2011). The dataset is composed by 13,626 university professors from Quebec (Canada) who have published at least one article indexed in the Web of Science (WoS) database during the 1980–2012 period. For every scholar in the dataset, different information has been collected, including their biological (BIRTH) and academic (PHD) ages, along with other bibliometric data, such as the year of first publication (YFP), number of publications in WoS (P), the proportion of publications with the scholar in the first position (PP_POS_FIRST), the proportion of publications with any type of international collaboration (PP_INT_COLLAB), etc. The full list of variables considered can be found in Table A1 of the Appendix A.

The data also include information about the research domain of the scholars. A total of nine disciplinary fields of activity of the scholars are considered, based on the 2000 revision of the U.S. Classification of Instructional Programs (CIP)[2] developed by the U.S. Department of Education's National Center for Education Statistics (NCES). The nine fields of activity, as well as the distribution of researchers among the fields can be seen in Table A2 in the Appendix A.

For the robustness of the results, we have selected researchers that are born after 1960 and have obtained their PhD degree since 1980. Moreover, since the last recorded PhD year is 2005, we have selected only the researchers that have their first publication the latest in 2010. Therefore the variable YFP is bounded at 2010 and the data truncated correspondingly.

Our final dataset comprises of 3574 researchers. Using this sample, we will make inferences about the researchers, in general, who represent our statistical population. We believe our sample is representative for researchers, in general. The external validation of our analyses, using another dataset, will be deferred to another manuscript.

The subsequent analysis is divided in two main parts. Firstly, we will perform an 'overall analysis', for all the selected researchers in the dataset, regardless their field of activity. We employ linear regression models for average birth and PhD years, as well as for all individual observations. Secondly, we are also interested in the particular characteristics of researchers in different fields and examine the potential disciplinary differences in the results. We therefore apply a similar analysis at the field level.

3. Overall analysis

We start our analysis by investigating the Spearman rank correlation among all variables in the study (see Table A1 in the Appendix A). The correlation matrix is depicted in Fig. 1. The correlation plot illustrates the correlations between BIRTH and PHD with other variables, and also brings insight into the correlations between the different independent variables. The age-related variables are well correlated among themselves. That is, birth (BIRTH) and PhD year (PHD) of researchers exhibit a strong correlation. Moreover, the year of first publication (YFP) is the only independent variable that presents a substantial correlation with these two age-related variables. Fig. 1 provides clear evidence to support the idea that YFP is the

[1] Despite its strong (and sometimes unintuitive) assumptions that are frequently violated in practice, linear regression modelling remains nevertheless the typical (first) approach in investigating the relationships between the variables of interest and covariates.

[2] The Classification of Instructional Programs (CIP) is developed by the U.S. Department of Education's National Centre for Education Statistics (NCES). More details can be found at: http://nces.ed.gov/pubs2002/cip2000/