Journal of Informetrics 11 (2017) 629–644

Contents lists available at ScienceDirect

Journal of Informetrics

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

Search for *evergreens* in science: A functional data analysis

Ruizhi Zhang [a], Jian Wang [b,c,*], Yajun Mei [a]

[a] H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology, 755 Ferst Drive, NW, Atlanta, GA 30332, United States

[b] Center for R&D Monitoring (ECOOM) and Department of Managerial Economics, Strategy and Innovation, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

[c] German Center for Higher Education Research and Science Studies, DZHW Berlin, Schuetzenstrasse 6a, 10117 Berlin, Germany

ARTICLE INFO

Article history:
Received 6 December 2016
Received in revised form 17 May 2017
Accepted 17 May 2017

Keywords:
Citation trajectory
Evergreen
Functional Poisson regression
Functional principal component analysis
K-means clustering

ABSTRACT

Evergreens in science are papers that display a continual rise in annual citations without decline, at least within a sufficiently long time period. Aiming to better understand *evergreens* in particular and patterns of citation trajectory in general, this paper develops a functional data analysis method to cluster citation trajectories of a sample of 1699 research papers published in 1980 in the American Physical Society (APS) journals. We propose a functional Poisson regression model for individual papers' citation trajectories, and fit the model to the observed 30-year citations of individual papers by functional principal component analysis and maximum likelihood estimation. Based on the estimated paper-specific coefficients, we apply the *K*-means clustering algorithm to cluster papers into different groups, for uncovering general types of citation trajectories. The result demonstrates the existence of an *evergreen* cluster of papers that do not exhibit any decline in annual citations over 30 years.

1. Introduction

Science is a skewed world where a small number of publications receive a disproportionate amount of citations. What do citation trajectories of the most cited papers look like? Do they follow the "typical" citation trajectory documented in the literature, specifically, the annual citations of a paper rise to a peak in the first few years after publication and then slowly fade away over time? Fig. 1 plots annual citations of the top ten most cited papers published in the American Physical Society (APS) journals, and their annual citations are counted in the Web of Science (WoS) from the year of publication to 2016. Among them the youngest was published in 1999, and the oldest 1964. Correspondingly, the length of their observed citation trajectories range from 18 to 53 years (Appendix A). In addition to their exceptionally large number of citations, a remarkable observation is that most of them (at least seven out of ten) do not even show any sign that their annual citations are about to peak and will start to decline in the near future. We refer to this phenomenon of continual rise in annual citations without decline as *evergreens*, which clearly violates the "typical" pattern of citation trajectory. Although we cannot predict whether these papers will remain highly cited in the future, the fact that they have not yet become obsolete after up to 53 years calls for attention, especially considering that the majority of papers reach their citation peak around the 3rd or 5th year after publication and that most bibliometric analyses examine citations in a relatively short time window.

* Corresponding author at: Naamsestraat 69, bus 3535, 3000 Leuven, Belgium.
E-mail address: jian.wang@kuleuven.be (J. Wang).

http://dx.doi.org/10.1016/j.joi.2017.05.007
1751-1577/

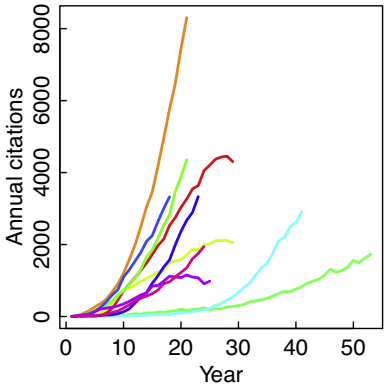

Fig. 1. Annual citations of the top ten cited APS papers. One curve represents one paper, and details about these ten papers are reported in Appendix A.

The objective of this paper is to better understand *evergreens* in particular and patterns of citation trajectory in general. Moreover, do *evergreens* constitute a general type of citation trajectory, or are they so rare that they cannot be captured in any statistical cluster analysis?

To this end, we develop a functional data analysis (FDA) method to analyze the 30-year citation trajectories of a sample of publications published in 1980 in APS journals. Our FDA method integrates functional principal components analysis, Poisson regression, and *K*-means clustering. More specifically, we model the citation trajectories of individual publications by a small number of common basis functions and paper-specific coefficients on these basis functions. For each paper, its 30-dimensional vector of citations can be characterized by its coefficients on the common basis functions, which subsequently serve as inputs for the *K*-means clustering, to uncover general types of citation trajectories. Results of our cluster analysis provide strong evidence that *evergreens* exist as a general class of citation trajectory. In addition, we are not able to predict whether a paper will become an *evergreen* by some *ex ante* paper features such as the number of authors and references.

The remainder of this paper is organized as follows. We begin with a brief review of previous cluster analyses of citation trajectories, as well as the method of functional data analysis, followed by a description of our dataset. Next, our proposed model and method is presented, with the emphasis on how to combine functional principal component analysis, Poisson regression, and *K*-means clustering algorithm for modeling and clustering citation trajectories. Then we report the empirical results of our proposed model and method to the real citation dataset. Implications of our findings are also discussed.

2. Prior literature

2.1. Clustering citation trajectories

Citation aging has been a long-standing research topic, and different patterns of citation trajectories have been documented in the bibliometrics literature (Aversa, 1985; Avramescu, 1979; Baumgartner & Leydesdorff, 2014; Garfield, 1980; Glänzel & Schoepflin, 1995; Line, 1993; Redner, 2005; Rogers, 2010; Wang, 2013). Aversa (1985) conducted probably the first rigorous statistical analysis of citation trajectories, investigating 9-year citation trajectories of 400 highly cited papers published in 1972 and applying the *K*-means clustering algorithm to the normalized annual citation counts (i.e., annual citations divided by total citations in the whole studied time period). Aversa (1985) identified two clusters: *delayed rise – slow decline* and *early rise – rapid decline*.

Costas, van Leeuwen, and van Raan (2010) analyzed about 30 million documents in WoS published between 1980 and 2008. Following Price's observation, documented in his personal communication to Aversa (1985), Costas et al. (2010) classified papers into three categories: 50% papers as *normal documents*, 25% as *delayed documents*, and 25% as *flashes-in-the-pan*. However, these three clusters are defined based on a single real-valued summary statistics of individual papers, *Year 50%*, defined as the year when a paper has cumulated half of its total citations up to year 2008. In addition, there is no statistical justification on the proportion of these three clusters.

More recently, Colavizza and Franceschet (2016) examined about half million papers published in APS journals and applied the spectral clustering method on the normalized annual citations received by these papers within the APS database. The three identified general types of citation trajectories are *middle-of-the-roads*, *sprinters*, and *marathoners*. *Middle-of-the-roads* papers display an average citation aging pattern, and can be viewed as corresponding to *normal documents*. *Sprinters* have an early and high peak and a fast decline, which can be viewed as *flashes-in-the-pan*. *Marathoners* represent "fast or slow-rise, moderately peaked histories, followed by a slow decline, or absence of decline, or even a constant rise in received citations over time" and therefore can correspond to *delayed documents* or *evergreens*.

The phenomenon of *evergreens*, which were emphasized by Avramescu (1979) and Price (see Aversa, 1985), were not identified by clustering analyses in Aversa (1985) and Costas et al. (2010), while *marathoners* in some specifications in Colavizza and Franceschet (2016) also display a continually increasing annual citation curve.