Journal of Informetrics 9 (2015) 449–454

Contents lists available at ScienceDirect

Journal of Informetrics

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

Compressing multiple scales of impact detection by Reference Publication Year Spectroscopy

Jordan A. Comins*, Thomas W. Hussey

Virginia Tech Applied Research Corporation, Arlington, VA, United States

ARTICLE INFO

Article history:
Received 20 January 2015
Received in revised form 24 February 2015
Accepted 20 March 2015
Available online 7 April 2015

Keywords:
Scientometrics
Informetrics
Citation analysis
Reference Publication Year Spectroscopy
Science policy
Funding agencies

ABSTRACT

Reference Publication Year Spectroscopy (RPYS) is a scientometric technique that effectively reveals punctuated peaks of historical scientific impact on a specified research field or technology. In many cases, a seminal discovery serves as the driving force underlying any given peak. Importantly, the results from a RPYS analyses are represented on their own distinct scales, the bounds of which vary considerably across analyses. This makes comparing years of punctuated impact across multiple RPYS analyses problematic. In this paper, we propose a data transformation and visualization technique that resolves this challenge. Specifically, using a rank-order normalization procedure, we compress the results of multiple RPYS analyses into a single, consistent rank scale that clearly highlights years of punctuated impact *across* RPYS analyses. We suggest that rank transformation increases the effectiveness of this scientometric technique to reveal the scope of historical impact of seminal works by allowing researchers to simultaneously consider results from multiple RPYS analyses.

1. Introduction

Citations signify the relevance of prior research or invention. In the scientific community, the aggregation of citations attributed to a specific work is commonly taken as a central indicator of its scholarly impact (De Solla Price, 1965; Garfield, Malin, & Small, 1978; Radicchi, Fortuno, & Castellano, 2008). More generally, citations and citation counts are thought to represent how knowledge accumulates, combines and transfers to generate new ideas and discoveries. Since citations function as linkages between scientific works, citation records provide an opportunity to quantitatively identify seminal contributions to a given research field or technology (Kostoff & Shlesinger, 2005; Marx, Bornmann, Barth, & Leydesdorff, 2014; van Raan, 2000).

One technique leveraging citations to detect important scientific contributions is Reference Publication Year Spectroscopy (RPYS). RPYS offers a quantitative approach to assist in identifying the historical roots of research fields and topics (Marx et al., 2014). To accomplish this, RPYS considers the references cited by a cohort of publications resulting from a particular database query. By way of example, consider a search query for topic X that returns only one article. If this article, published in 1990, cites a reference published in 1980, then 1980 is used as a data point in the foregoing analysis. As such, after a set of publications is returned from a database query, the publication date of each cited reference from this set of publications is extracted and mapped onto a frequency distribution sorted by time. The resulting visualization often reveals punctuated

* Corresponding author at: Virginia Tech Applied Research Corporation, Arlington, VA 22203, United States. Tel.: +1 703 879 8142.
E-mail address: jcomins@gmail.com (J.A. Comins).

http://dx.doi.org/10.1016/j.joi.2015.03.003
1751-1577/

Table 1
Search terms and the corresponding number of results yielded from the Web of Science Core Collection (1974-present).

WoS core collection topic search	No. of results (total)	No. of results (articles and conference proceedings only)	No. of cited references obtained
"Viterbi"	4840	4713	62,405
"convolutional code" OR "convolutional codes"	4011	3863	59,586
"hiddenmarkov model" OR "hidden markov models"	2348	2312	58,648
"continuous speech recognition"	1569	1538	25,070
"automatic speech recognition"	3427	3370	66,233
"speech recognition"	17,946	17,462	330,250

peaks in the distribution. These peaks correspond to years containing a larger number of cited references within discrete bins of time. Often, these peaks are driven by a large number of references to a seminal work in the field. To date, RPYS has been successfully applied to investigations of important early contributions in several research topics (Leydesdorff, Bornmann, Marx, & Milojević, 2014; Marx et al., 2014; Marx & Bornmann, 2013; Wray & Bornmann, 2014).

There is, however, a major challenge with the current methodology. Namely, the results produced by a given RPYS analysis are represented within their own distinct range or scale, the bounds of which vary considerably across analyses. In other words, using the presently defined RPYS technique, it could be difficult to compare patterns of maxima for the cited references of a small research fields with those of the cited references from a much larger research fields. Making RPYS analyses amenable to large-scale comparative analysis is an important extension of the technique for future applications. For instance, it would allow analysts to more readily evaluate whether a large number of research topics show a similar history of important findings as demonstrated through cited works. A second possibility is that being able to estimate the similarities between the citation histories for various subfields might open-up an entirely novel venue for defining the relationships between these subfields – with the assumption being that research areas correspondingly informed by the same seminal works are more similar than those that are not.

To address this shortcoming, we demonstrate here how the addition of a simple data transformation procedure to the standard RPYS methodology can aid in the detection of shared patterns of maxima for the cited references across RPYS analyses, which potentially suggest common historical influences. Specifically, we adopt the use of a rank-transformation procedure commonly used in inferential statistics to perform non-parametric analyses (Conover & Iman, 1981; Labovitz, 1970). This transformation compresses the multiple scales produced from various RPYS analyses into a single rank scale that allows researchers to identify years of punctuated impact *across* RPYS analyses. We describe a visualization procedure that efficiently represents data from multiple RPYS analyses concurrently.

To demonstrate the efficacy of this procedure, we begin with a publication that we suspect a priori has meaningfully impacted numerous research topics: the Viterbi algorithm first published by Andrew Viterbi in 1967. In this groundbreaking work, Viterbi describes an algorithm that identifies the most likely sequence of hidden states associated with a sequence of known or observed states. The algorithm is widely used in stochastic models and error-correcting (or decoding convolutional codes) as well as in a variety of computational procedures pertaining to machine speech recognition (Viterbi, 2006). Given this, we performed six RPYS analyses for research topics pulled from the Web of Science that all pertained to the development or use of the Viterbi algorithm (Viterbi, 1967). The Viterbi algorithm's impact and use in a wide array of research communities, from statisticians studying stochastic models to engineers and computer scientists working on various aspects of machine speech recognition, make it an ideal candidate for demonstrating the value of this data-transformation procedure for comparing multiple RPYS analyses concurrently.

2. Method

We accessed and downloaded data from the Thomson Reuters Web of Science (WoS) between December 12, 2014 and December 14, 2014. We performed topic searches using the Web of Science Core Collection, for which we had back-records from 1974 to 2014. The topic searches performed for our six RPYS analyses were as follows: (1) "Viterbi", (2) "convolutional code" OR "convolutional codes", (3) "hiddenmarkov model" OR "hidden markov models", (4) "continuous speech recognition", (5) "automatic speech recognition" and, finally, (6) "speech recognition". The number of results for these six searches is shown in Table 1. We then filtered these results to only represent articles and conference proceedings (again, values in Table 1). The last step of the data collection process was to download the results from WoS Core Collection using the "Save to Other File Formats" option and selecting "Full Record and Cited References" as our desired record content.

To extract cited references from the data we obtained, we used the open source Sci2 tool (Sci2 Team, 2009) developed by Indiana University. This converts the Full ISI records into tables. This conversion conveniently represents "Cited References" as a single column within this table. The resulting data contained in this column uses a | character as a delimiter between the different cited references. Using a python script, we split our cited references data using this delimiter to generate a new table representing all of the cited references. This process allows us to effectively isolate properly structured references and their associated publication year from WoS records.

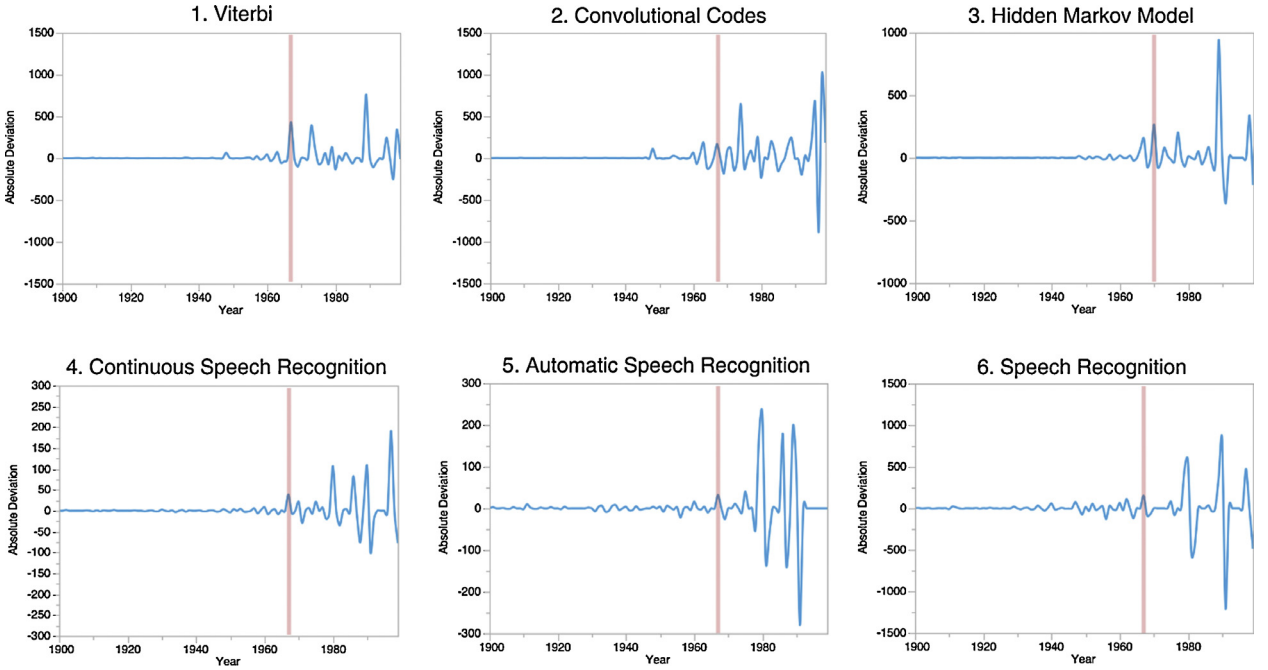

Fig. 1. RPYS results for six separate topic searches. Here we represent our results in a canonical fashion for RPYS analyses by plotting the absolute deviation of cited references for a given year from the median of a 5-year epoch. Within any individual RPYS analysis, this accentuates when peaks occurred.

3. Results

Results from RPYS analyses are most typically represented in two ways: (1) the raw number of cited references occurring per year or (2) the absolute deviation of the number of cited references on a particular year from the 5-year median. The second method of visualization is often favored due to the fact that the raw number of cited references increases toward the present date. By taking the absolute median cited references from a 5-year median, one makes it easier to identify works that might be only modestly cited overall but were very influential in the context of when they were published. Thus, Fig. 1 shows the RPYS analyses for our six topic searches in canonical absolute deviation form. Each analysis reveals multiple maxima; each result also yields a peak occurring in 1967, the year Andrew Viterbi published his seminal technique later dubbed the Viterbi algorithm (year highlighted in year). In all six cases, the most cited reference from 1967 was indeed the Viterbi algorithm.

Importantly, the range of values of the maxima across the RPYS varies considerably. For instance, the largest value of absolute deviation in a 5-year period for our search on convolutional codes was 1010, as opposed to 191 for our search on continuous speech recognition. Thus, despite the fact that 1967 appears as a punctuated peak of historical scientific impact for each of these research topics, overlaying these plots makes the task of identifying any year of mutual importance difficult. This challenge is demonstrated in Fig. 2.

To overcome the difficulty in making cross RPYS comparisons, we rank-transformed the deviation data shown in Fig. 1. Before detailing the ranking procedure, however, we note the importance of applying the transformation method to the deviation data. This is because the deviation procedure provides the necessary temporal control to allow important early works to stand out in an RPYS analysis. As has been noted in prior work using RPYS, references tend to point to more recent publications, as older discoveries either are incorporated into multiple sources (i.e., textbooks) to become familiar knowledge within a community or are associated with a more recent author (Leydesdorff et al., 2014; Marx et al., 2014). One helps control for this tendency with the deviation procedure, which permits seminal early works to emerge and be used for making interesting comparisons about the shared seminal works across research fields.

To rank transform data, one takes the complete set of n values from a given RPYS analyses, say $X_1, X_2, \ldots, X_n$. These observations are then sorted by the magnitude of the observed values. These values are then substituted in such a manner that the largest value takes the value n, the second largest take the value $n-1$, and so on.

Our rationale for selecting a rank transformation procedure was based on both the attributes of the deviation data itself as well as limiting our transformation method to one found in common statistical packages. Regarding the attributes of the data, RPYS deviation values are signed (i.e., positive or negative) integers. As a result, these values are not compatible with several common transformations for the following reasons:

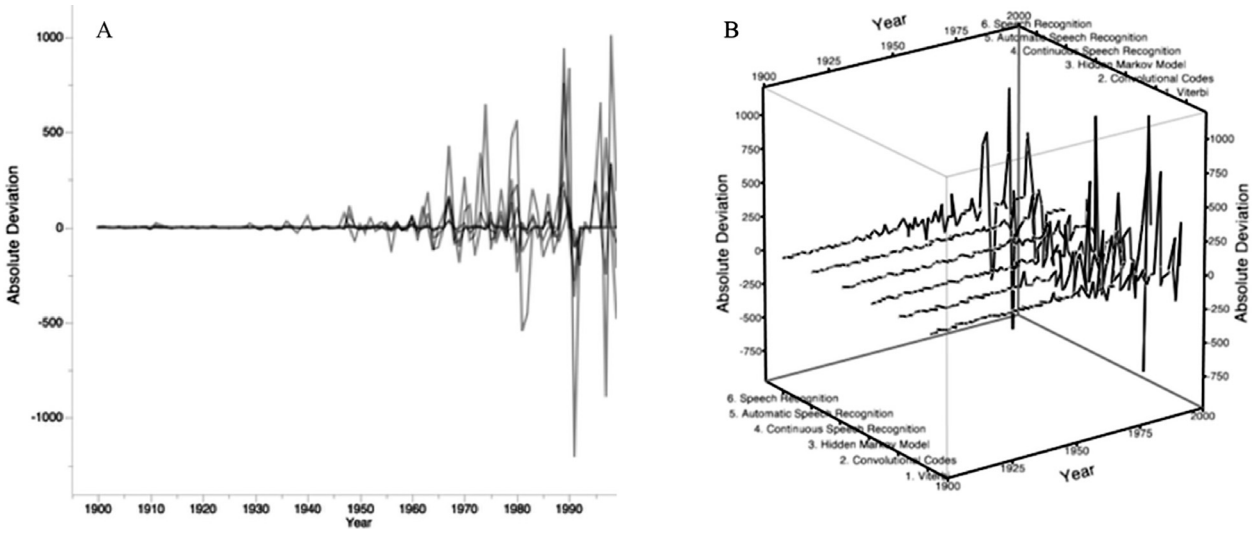

Fig. 2. Simply overlaying two- (A) or three-dimensional (B) representations of RPYS results across separate topic searches makes it difficult to identify patterns of maxima occurring *across* the individual RPYS analyses.

- *Logarithmic Transform*: Log transforming the data requires the values to be both non-negative and non-zero. For this procedure to be used, one must either omit all negatives values and zeroes or one would need to add a constant k to each resulting RPYS deviation value to make the minimum value of the results greater than 0. This k would vary across RPYS analyses.
- *Power Transform*: Power transforms, such as the common square root transform, requires the values to be non-negative. For this procedure to be used one must either omit negative values or one would need to add a constant k to each resulting RPYS deviation value to make all values non-negative. This k would vary across RPYS analyses.
- *Arcsine Transform*: Arcsine transforms require the data be bounded [0, 1].
- *Logit Transform*: Logit transforms require the data be bounded (0, 1).

Thus, rank transforming presents a simple, common method for transforming the deviation data that does not require individuals to devise with a custom transformation procedure per RPYS analysis.

Beyond their ease of use and applicability to RPYS data, rank transformations offer two distinct advantages: (1) it compresses the multiple scales of RPYS analyses into a single, consistent rank scale and (2) it allows researchers an intuitive way to filter and visualize cross RPYS comparisons. Regarding this latter point, by rank transforming we can easily set an arbitrary threshold to scan for years containing a certain degree of cited references. For example, for each of our RPYS analyses, we obtained 100 data points. By using a rank transform, we can easily compare the top n years with the highest RPY deviations across our analyses.

The advantages of the rank-transform are shown in Fig. 3. Whereas presentation in Fig. 1 graphs the deviation results of multiple RPYS analyses separately, the heatmap in Fig. 3 visualizes multiple RPYS analyses concurrently. And, while here we compared a modest number of RPYS analyses, our technique generalizes so that one could perform and visualize dozens of RPYS analyses simultaneously. In such cases, visualizing the data as they are in Fig. 3 is one method to more readily compare multiple RPYS results.

Fig. 3a plots the data from Fig. 1 converted into a heatmap. As the scales of the six RPYS analyses varied considerably, identifying patterns across analyses using this heatmap is extremely difficult. The heatmap in Fig. 3b, however, has converted each deviation value to a rank and set a threshold for the midpoint of the color map at rank 90 (thereby highlighting the 10 years with the greatest deviations, or largest detectable impact, per analysis; we also note that similar heatmaps can be created from WoS data using the freeware RPYS.EXE; Marx et al., 2014). As a final note, we add that there are multiple methods and motivations for visualizing RPYS results. And, though the heatmap is one means for efficient cross RPYS comparison, when one is solely interested in a single RPYS result they should continue to use the untransformed data as shown in Fig. 1.

Using our procedure, we find a salient pattern emerging across analyses at the 1967 band, which corresponds to the year the Viterbi algorithm was published (further scrutiny indeed verifies the Viterbi paper underlies this pattern). Our results show that this year contained an exceptionally high number of cited references for each of the six RPYS analyses, validating our methodology for cross RPYS analyses comparison.

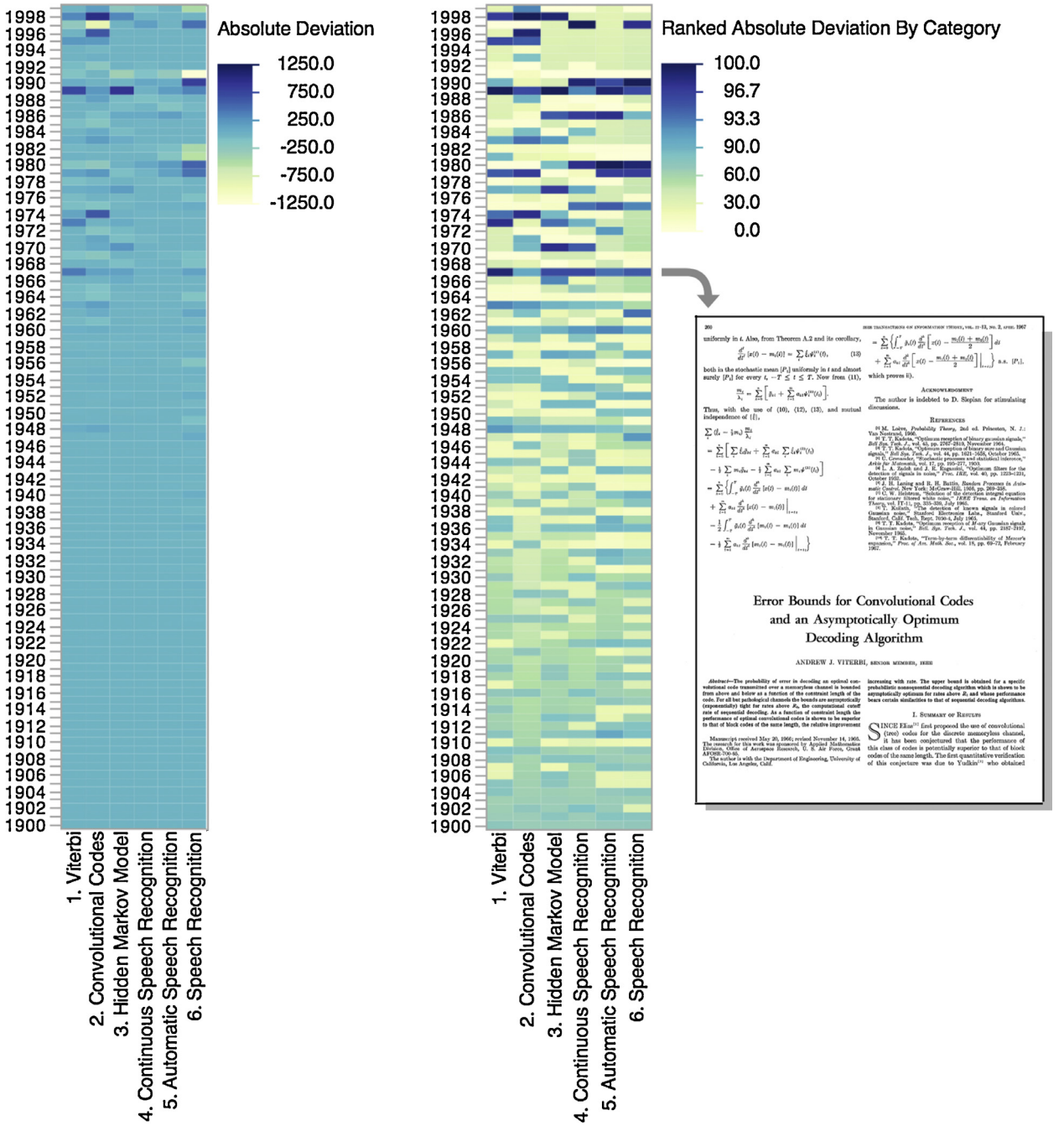

Fig. 3. Heatmaps representing six separate RPYS analyses. The gradient in the heatmap on the left relates to the absolute deviation of cited references shown in Fig. 1. It is difficult to observe patterns across RPYS analyses using this schema. The gradient in the heatmap on the right, however, represents rank transformed absolute deviation of cited references. We can further set the gradient scale to emphasize cells with the largest deviation in cited references (ranks 90–100). The resulting visualization readily highlights 1967, the year the Viterbi algorithm was published.

4. Discussion

In this short report, we consider the challenge of comparing results from multiple RPYS analyses. Due to the fact that the size of different research topics or areas can vary substantially, this makes the task of identifying similarities across individual RPYS analyses problematic. We show that a simple rank-transform of the resulting vector of the absolute deviation of cited references within RPYS analyses makes them amenable to comparative evaluation. Specifically, the transformation provides

two distinct advantages: (1) it collapses the results from multiple RPYS analyses into a single scale and (2) this rank order scale is highly amenable to filtering techniques to visualize the *n* years containing the most cited references.

Making RPYS analyses conducive for comparative study offers several important future directions for scientometric research. First, one can now more readily evaluate whether different topics or keywords of research areas have a shared history of important findings as demonstrated through cited works. And, second, being able to estimate the similarities between such citation histories for fields might allow researchers another venue for defining the relationships between subfields and topics together. In other words, perhaps research areas similarly informed by the same seminal works are more similar than those that are not.

Conflict of interest

JAC works for the non-profit Virginia Tech Applied Research Corporation (VT-ARC), which supports the Air Force Office of Scientific Research (AFOSR). TWH consults for VT-ARC and is the former Chief Scientist of AFOSR. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory.

Acknowledgements

Effort sponsored in whole or in part by the Air Force Research Laboratory, USAF, under Partnership Intermediary No. FA9550-13-3-0001. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

The authors thank Stephanie Carmack and two anonymous reviewers for constructive feedback.

References

Conover, W., & Iman, R. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, *35*(3), 124–129. Retrieved from http://amstat.tandfonline.com/doi/abs/10.1080/00031305.1981.10479327

De Solla Price, D. J. (1965). Networks of scientific papers. *Science*, *149*(3683), 510–515. Retrieved from http://www.eigenfactor.org/methods.php/methods.pdf

Garfield, E., Malin, M., & Small, H. (1978). Citation data as science indicators. In *Toward a metric of science: The advent of science indicators*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.15.4233

Kostoff, R., & Shlesinger, M. (2005). CAB: Citation-assisted background. *Scientometrics*, *62*, 199–212. Retrieved from http://www.akademiai.com/index/R56K1LR0H3605886.pdf

Labovitz, S. (1970). The assignment of numbers to rank order categories. *American Sociological Review*, *35*, 515–524. Retrieved from http://www.jstor.org/stable/2092993

Leydesdorff, L., Bornmann, L., Marx, W., & Milojević, S. (2014). Referenced Publication Years Spectroscopy applied to iMetrics: Scientometrics, Journal of Informetrics, and a relevant subset of JASIST. *Journal of Informetrics*, *1901*, 1–34. Retrieved from http://www.sciencedirect.com/science/article/pii/S1751157713001077

Marx, W., & Bornmann, L. (2013). Tracing the origin of a scientific legend by reference publication year spectroscopy (RPYS): The legend of the Darwin finches. *Scientometrics*, *99*(3), 839–844. http://dx.doi.org/10.1007/s11192-013-1200-8

Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology*, *65*(4), 751–764. http://dx.doi.org/10.1002/asi.23089

Radicchi, F., Fortuno, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(45), 17268–17272. Retrieved from http://www.pnas.org/content/105/45/17268.short

Sci2 Team. (2009). *Science of science (Sci2) tool*. Indiana University and SciTech Strategies. https://sci2.cns.iu.edu

van Raan, A. (2000). On growth, ageing, and fractal differentiation of science. *Scientometrics*, *47*, 347–362. Retrieved from http://link.springer.com/article/10.1007/s11192-014-1465-6

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1054010

Viterbi, A. (2006). A personal history of the Viterbi algorithm. *IEEE Signal Processing Magazine*, 120–142. Retrieved from http://www.fceia.unr.edu.ar/prodivoz/Viterbi_2006.pdf

Wray, K., & Bornmann, L. (2014). Philosophy of science viewed through the lense of "Referenced Publication Years Spectroscopy" (RPYS). *Scientometrics*, 1–19. Retrieved from http://link.springer.com/article/10.1007/s11192-014-1465-6