Telling the whole story: finding structures in bibliometric information using PCA

DOI:
10.13140/RG.2.2.25764.07041

Document Version
Other version

Telling the whole story: finding structures in bibliometric information using PCA

Working Paper · March 2017

DOI: 10.13140/RG.2.2.25764.07041

2 authors, including:

John Rigby

The University of Manchester

49 PUBLICATIONS **244** CITATIONS

SEE PROFILE

Telling the whole story: finding structures in bibliometric information using PCA

Keith Julian[1] and John Rigby[2]

[1]*Keith.Julian@manchester.ac.uk*

Manchester Institute of Innovation Research, Manchester Business School, Oxford Road, Manchester, United Kingdom, M13 9PL

[2] *John.Rigby@manchester.ac.uk*
Manchester Institute of Innovation Research, Manchester Business School, Oxford Road, Manchester, United Kingdom, M13 9PL (corresponding author) +44 (0)161 275 5928
Orcid ID: 0000-0001-9833-5965

Abstract

As bibliometric data is multidimensional, its study, by means of index numbers, especially index numbers in the form of ratios, rarely captures all the information. Across the bibliometric and evaluation literatures there is increasing scepticism about the value that single index numbers give to the understanding of scientific behaviour and its consequences including impact. The authors propose and demonstrate the use of a multivariate approach – principal component analysis - that gives greater insight into the aspects of scientific publication. Principal component analysis is put forward as the most suitable multivariate method available as it does not emphasise any variable and identifies important signals in the data that may not be observed with univariate methods. A data set analysed in a previous piece of work on double-dipping is used here and is subject to PCA which reveals three components, an input related component, an output related component and an outcome (impact) related one. Importantly, citation is shown to be of limited significance in explaining overall variability within the data set.

Keywords

Indicators; science policy; methods and techniques; principal components analysis

Introduction

Science studies and evaluation have developed considerable insight through bibliometric methods into the structure and working of scientific practice, knowledge production systems and, and impact. Common to many of these various approaches is the use of index measures, such as the journal impact factor and more recently, the Hirsch index. Index numbers have considerable currency in research policy and management, being seen as relatively unproblematic benchmarks that, if sensibly used, can steer and manage science, its institutions and practitioners effectively. But while many remain invested in the view that index numbers are efficient, there is also the view that such measures are “economical with the truth” (Oxford Dictionary of Modern Quotations, 2007), and the growing unease about their

perverse effects has led to various public expressions of disquiet (The American Society for Cell Biology (ASCB), 2013).

An alternative approach to understanding science and its institutions is to employ statistical methods that draw on a greater range of bibliometric data types. We outline the need for and relevance of one of these multivariate methods, Principal Components Analysis (PCA), which we then demonstrate on a data set of publications. Our aim in this paper is to encourage a move in this direction and to give a more systematic understanding of the key features of science and its institutions by adopting a statistical approach that uses multi-variate methods, in this case principal component analysis (PCA). This is an approach that avoids the narrow focus on one single dimension of bibliometric data that is very characteristic of many studies where citation is the dependent variable that is related to other factors (Baldi, 1998; Lewison & Dawson, 1998). Here we seek to put the citation feature of the data into the overall context, and reveal through the use of PCA methods some of the key dimensions and important features of scientific activity.

Literature

Despite the awareness of the many and complex dimensions of bibliometric data, a consistent and regular flow of index and ratio measures has emerged to exploit the quantitative properties of data, the best-known being the journal impact factor and more recently, the h-index. Such indicators are often then criticised, or shown to be flawed or inadequate and variations or new ones are introduced in the attempt to improve upon them. Single index numbers have the apparent benefit of simplicity. Indeed, they have the outward appearance of being comparable one with another as the impression can be given that some standardization or normalization has occurred in their creation; but in practice, changes in the value of an indicator may come either from the numerator or the denominator of the index ratio, making comparison of these indices as they relate to individual papers or institutions, or research fields etc., difficult and potentially misleading. The use of the journal impact factor, a pre-

eminent index metric, has now begun to draw widespread criticism and create concern amongst a range of individuals and organisations.

This concern has been in evidence in many places, for example amongst Nobel Prize Winners, see (Sample, 2013) on Randy Schekman's comments on receiving his prize for physiology in 2013, but also amongst professional scientific organisations, for example (Setti, 2013) on the IEEE. Amongst researchers themselves there has been consternation at how a single paper can affect the impact factor of a journal (Dimitrov, Kaveri, & Bayry, 2010).

The problem lies with the single indicator approaches to what are complex systems. The controversy surrounding the h-index and other one dimensional metrics (Gingras, 2014) provides support for approaches that attempt to understand science as irrevocably multi-faceted science system. To some degree, the initial concerns of bibliometric research – to understand how scientific knowledge is created and how the institutions of science function – have been joined by diverse attempts to develop bibliometric tools for very specific policy purposes. This development has been supported in part by those in science studies themselves, although there is a realization that a more holistic approach is now needed.

A univariate statistical approach cannot by definition describe in a comprehensive manner the many characteristics, the measurements and inter-relationships of the scientific research reported by a paper in a journal, from an institution, supported by a funding body, or in a particular field. Univariate methods are incomplete because they generate separate metrics that do not take account of correlations between them. Principal component analysis (PCA) is the most suitable multivariate method available for the task of assessing the common interrelations of variables as regards the whole as it does not emphasise any single variable. By contrast, multivariate regression using the count of citations as a dependent variable would emphasise the primacy of that measure as earlier univariate studies do, while factor analysis requires the proposal of a formal model for the analysis.

The use of citation counts in a univariate analysis to infer the quality of scientific research is therefore, in our view, akin to using height alone to determine the health of a human population. If we wish to look at concepts like health through height alone we will not be using data about weight, diet, age, income, environmental exposure etc. which are all known to be important with regard to health. In health studies there is a measure "Body Mass Index - BMI" which uses height and weight, instead of relying on height alone. The BMI index for health can be compared to the Hirsch and other indices in bibliometrics. However the medical literature recognizes that this single index number is vastly complicated measure (Davey Smith et al., 2009) and inadequate by itself as an indicator and or predictor of health outcomes. Its use is frequently accompanied in health advice documents by the use of 2 - dimensional graphs of height/weight for men and for women, each with marked regions of risk of a third variable such as obesity or heart problems. For the same reason that health metrics seek to describe a complex underlying reality, the authors propose that a similar multidimensional analysis should be applied to assess the "condition" of papers, institutions, and research fields.

The application of multivariate methods to the study of bibliometric data has a long history and some uses of PCA have already been made. Studies using web links as analogous to citation and as a means of identifying web communities and networks by (Faba-Perez, Guerrero-Bote, & De Moya-Anegon, 2003, 2004) are a good example of attempts to reveal structure in the use of the Internet. PCA has also been used as a way of assessing the impacts of papers, contributing, to some extent, to the debate on the specific aspects of impact (Bornmann, 2015) and its dimensionality (Chen, Tang, Wang, & Hsiang, 2015), see also Bollen, Van de Sompel, Hagberg, and Chute (2009) and their detailed study of impact measures. Hendrix (2008) has also used PCA to focus on research production in the US health research system and has revealed some important aspects of the structure of knowledge generation in relation to a particular group of knowledge producing organisations, i.e. medical schools. However, to our knowledge, these studies have not sought to identify structural features of publication in the broader sense using all the available data types from citation indexes.

The authors recently published a study of the citations generated by papers funded by HFSP and EMBO, or both, to discover information about "double-dipping". The data set contained information about ten characteristics of the papers, but only one measure at a time was used in the analysis by a univariate method. We therefore considered that it would be revealing to re-subject this data set multivariate methods and to exploit the full set of data from ISI which contained information about other characteristics which were measured by scalar units that had not been used in the analysis. It does not matter that there is no known relationship between these characteristics if we do not want to form an index number from these counts; a multidimensional representation can be used to study the correlations, find what weight they have, and compare them. An example of information that had previously not been incorporated in the analysis is that of the page length of individual papers. As journals have a relatively inelastic count of pages each year, and therefore any single paper is in competition with others, this results in a competition for space. Paper length and count of papers are controlled by a publisher, and this may be reflected as a restriction on the opportunity for citations and therefore place a restriction on any calculated index.

All scalar measurements of characteristics can be used to increase the dimensions of analysis in bibliometrics. Multivariate analysis by Principal Component analysis, Factor analysis, Discriminant analysis and their equivalent graphical representations have been used in medicine, psychology and other fields over many years. Introductory texts are by Manly (1986), Cooper and Weekes (1983) and Fisher and van Belle (1993) ENREF 10, the latter describing the statistical calculations for PCA with examples from the medical field. There are many later, more mathematical, texts available. It is useful to consult those that are cited by the writers of the computer programs used in the data manipulation, such as SPSS or SYSTAT, which is used in this paper.

Method and Results

Data set

The data set used here is the same set of data employed in the paper (Rigby & Julian, 2014). In this data set, papers were selected from the Web of Science (WOS) for the period 2008 2013 inclusive and a range of data was downloaded and reviewed. Papers were chosen for this period as they were first to contain systematic data on funding acknowledgements. Papers were selected on the basis that they had funding from either the Human Frontier Science Program and or the European Molecular Biology Organisation. These papers represent the outputs of research conducted by leading scientists and supported by funding bodies (funding organisations) of high repute, whose grants provide resource for research in the area of molecular biology. In this analysis undertaken for this paper however, we focused upon the count of funding acknowledgements for each paper rather than making a distinction between these two major funding bodies, and upon other variables, in order to elicit more information (in the form of principal components) from the data.

In all, there were 5775 papers in the data set, with 5147 articles and 628 review papers. As the characteristics of articles and review papers are different, these were analysed separately. The data had been reviewed in Vantage Point in order to disambiguate funding body information.

PCA Method

The variables available to us from the data set we collected for our earlier study are 7 in number and are as in the following table, Table 1.

Table 1 Variables in the Data

Category or Variable	Code
Log(10) (1+ Times Cited) Log10(1+TC)	AA
Count of Authors	E
Count of Cited References	F
Count of Pages	G
Count of Countries	H
Count of Author Affiliations	J
Count of Funding Acknowledgements	K

The introductory texts cited above have geometric explanations which are most helpful to those who are not familiar with matrix algebra methods. In brief, with PCA, the n-dimensional geometrical space occupied by the individual scalar measurements is combined, rotated and projected to a lower dimension space with fewer vector components arranged to be as independent (orthogonal) as possible. In the literature this is sometimes called "projection to latent structures".

Computer-based PCA programs do not reduce automatically the dimensionality of the data, or project to latent structures, and there are choices that have to be made during the analysis. One choice is whether to use Correlation or Covariance matrix methods, and another the number of Varimax, or other rotations. In this paper, Correlation methods and a Varimax rotation were used. (For explanations of what these are the reader is referred to standard texts and in particular the handbooks of the relevant computer programs because the defaults can vary across programs). The standard correlation PCA methods, reflected in the computer programs available to us, make a single, common, type of transformation or standardisation on the variables before calculations are made. The mean value of each variable is subtracted from each original value and the result divided by the standard deviation of that set of variable values. This usually has the benefit that the numerical range of the standardised individual factors in any PCA study is comparable, and graphical representations are not dominated by one component. In the case of the variable "Times cited", the original data is highly skewed. We therefore choose to use the Log10(x+1) transformation before carrying out the

standardisation in the PCA calculations. (In the earlier paper, the authors used non-parametric statistics to avoid the problem).

Seven original variables were used in the analysis and seven new variables (also called components) automatically resulted. The full details of the calculations made with the Systat and SPSS computer programs are not given in the results below, only the significant features. By convention, a selection of the Principal Components is made to a cut-off point of 1.0 in the Eigenvalues, and/or a cumulative variance of approximately 70%.

Analysis of Articles and Review Papers

In the case of the articles data set, the first three Principal Components were chosen to the 0.984 level with a cumulative variance of 72%. The behaviour of the Eigenvalues of all the components (factors) is shown in the cumulative plot, known as a scree plot and this is shown in Figure 1. The first three principal components of the review articles were chosen with a cumulative variance of 79%. The scree plot for the review articles is shown as Figure 2. Broadly, both articles and review papers have similar latent structures with a three component solution.

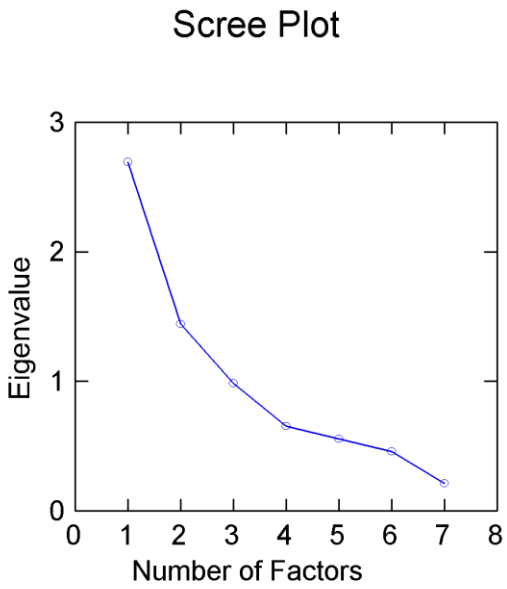

Figure 1 Scree Plot, Articles

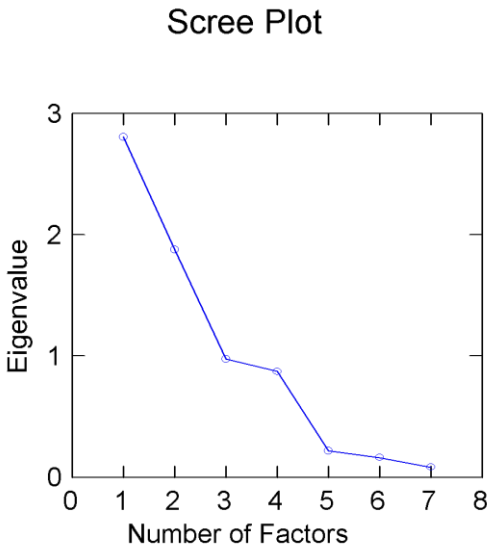

Figure 2 Scree Plot, Reviews

Results of the Correlation and Varimax calculations for Loadings and Variance

The graphical representation of the rotated Varimax results is given in the Component (Factor) loadings for both articles and review papers respectively. These are shown below in Figure 3 and Figure 4. PCA has produced three new uncorrelated orthogonal Components (Factors) and the individual variable loadings vectors form three groups which are nearly orthogonal, and this feature is discussed below.

	1	2	3
AA	0.085	-0.019	0.985
E	0.852	-0.027	0.128
F	-0.002	0.856	0.012
G	0.087	0.842	-0.029
H	0.803	0.033	-0.093
J	0.908	0.011	0.010
K	0.658	0.123	0.141

Table 2 (Component) Rotated Factor Loading Values: Articles

	1	2	3	
AA	-0.027	0.013	0.935	
E	0.943	0.027	0.093	
F	-0.041	0.951	0.098	
G	0.037	0.955	0.048	
H	0.919	-0.031	0.081	
J	0.954	-0.011	0.095	
K	0.290	0.135	0.429	

Table 3 (Component) Rotated Factor Loading Values: Review Papers

Variance Explained by Components			
	1	2	3
	2.693	1.443	0.984
Percent of Total Variance Explained			
	1	2	3
	38.465	20.615	14.058

Table 4 Variances Explained: Articles

Variance Explained by Components			
	1	2	3
	2.730	1.836	1.095
Percent of Total Variance Explained			
	1	2	3
	38.998	26.226	15.643

Table 5 Variances Explained: Review Papers

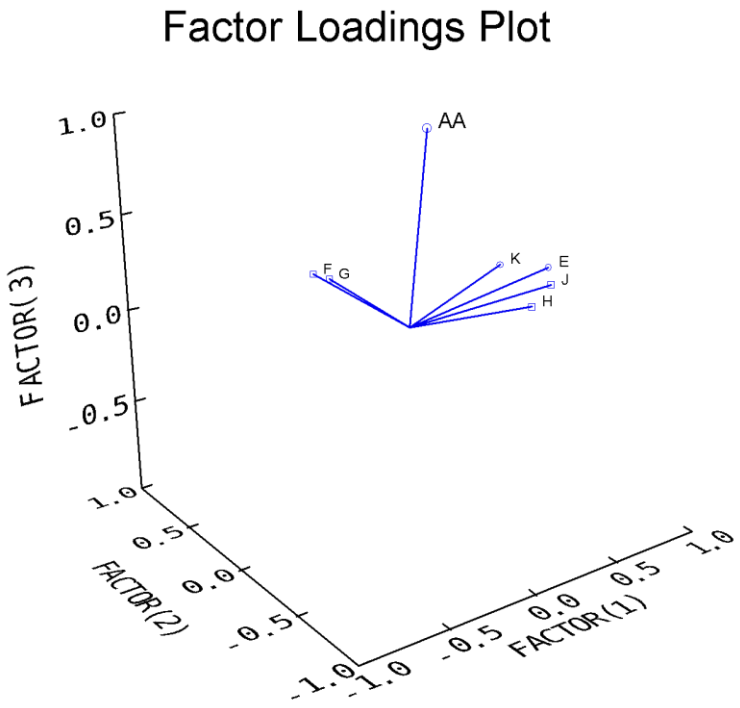

Figure 3 Factor Loadings Plot: Articles

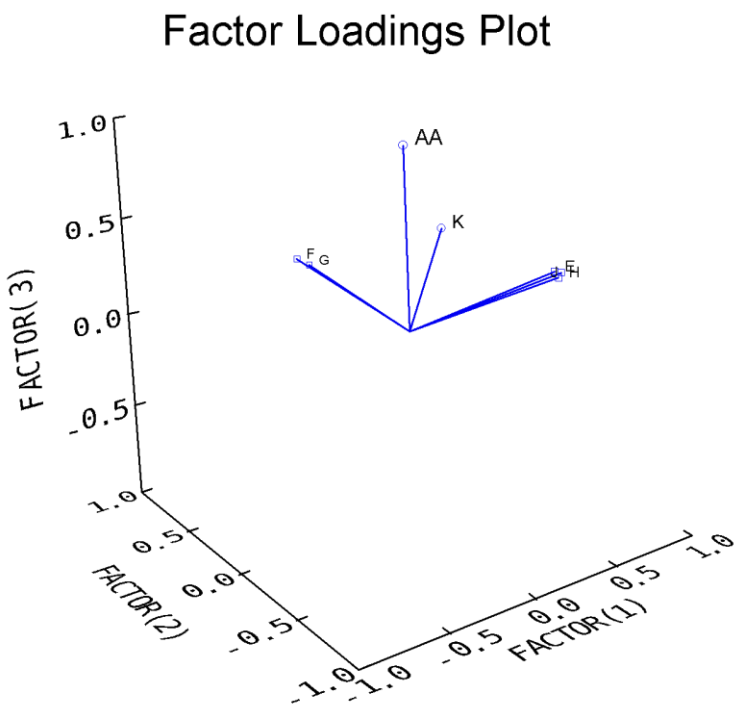

Figure 4 Factor Loadings Plot: Review Papers

Interpretations of the Component (Factor) loadings:

The component loadings of Component 1 for both articles and for the review papers include the count of authors, the count of countries, the count of authors' affiliations and the count of funding acknowledgements. Amongst the review papers however, the count of funding acknowledgements is considered less important, loading less on this first component. Our observation of this component is that the characteristics of these four variables considered together can be summarised by the description "Input". Together they describe the resources that were put into the research at the beginning.

The component loadings of Component 2 have similar values for the articles and the review papers. The variables which load on component two are the cited reference count and the number of pages in the published paper. We consider that the characteristics of these two variables considered together could be seen as the "Output" dimension or "component" that is present in the paper as published and revealed by the quantitative data we possess from the bibliographic record. We also note that for articles, this component is less important in terms of the overall explanation of variability than for review papers. Review papers are likely to cite more references.

Moving to the remaining variables which load on component 3 for the articles and for the review papers, we see that the cited reference count is by far the most important, and loads similarly. This component describes the "Outcome", the response of the world after the paper has been published. (To avoid confusion we avoid calling this Component "Impact").

Discussion

In the context of the use of "Times Cited", or its derivatives, as a leading measurement in current bibliometrics the most notable feature of the analysis we present here is that the third component for articles and review papers is the most heavily loaded by this measure but it capture only approximately 15% of the variation in the multivariate data set (for articles, the value is 14.52, while for review papers it is 15.64). When "Times Cited" (c.f. our "Outcome") is used as the only measure of the "character and importance within the bibliographic record", the more important features of published papers expressed in Components 1 and 2 for articles and review papers are being ignored. It is not surprising that concern is being expressed about the use of "Times Cited" alone, this analysis shows that this metric captures only a very small number of the actual measurable differences that exist between papers.

Earlier in this paper we suggested that the editorial publication process may be an important feature of research and publication. We believe that this is confirmed by the presence of an Output component formed from the cited reference count, (the count of citations made in a

paper by the authors), and the count of pages used or allowed to the authors, which captures 20% and 26% the variation (for articles and review papers respectively). The common behaviour of authors in choosing carefully where to submit their paper on the basis of perceived editorial policies and actions appears to be justified by the outcome they receive.

Some years ago (Lewison & Dawson, 1998) suggested that the count of funding acknowledgements of a paper was generally a predictor of higher impact and therefore a desirable outcome, an association that held because research with more funding had more successful peer review and was therefore more likely to be of higher quality. The insight of Lewison and Dawson about funding acknowledgements can be seen from the weight given to "Count of funding organisations" in the first component. However their explanation is incomplete as it does not take account of the other larger, contributing variables which also load on component 1, which are the counts of author affiliations, the counts of authors and the counts of countries involved in the publication of the research. In our analysis, both for article and for review papers, Component 1 captures approximately 38% of the variability in our data set, approximately the same as next two Components combined.

Conclusion

In studies of the benefits of international cooperation and interdisciplinary research the "Times Cited" or one of its derivatives is often used as a primary measure of the character of the research publication process. This study suggests that that approach is partial as it captures only a low proportion of the variability in any data set and limits our understanding of the essential characteristics of scientific knowledge generation. By processing all available variables, a more informed and fuller picture results with, as our analysis has shown, three key components to research activity – Input, Output, and Outcome. For example, the peer review process is said to begin before papers are submitted to journals, and is more rigorous or more filtered in international, interdisciplinary work. In this study the input variables in Component 1 show the opportunity for this behaviour is important, but it does not say it improves the quality of the work if this is measured by Component 3. As noted above, the

variables loadings in these two Components are very nearly orthogonal and have a low correlation.

This work shows the reason for the results in an earlier paper (Rigby & Julian, 2014) . We re-used this data here to make the comparison of a multivariate PCA with the previous univariate investigation by "Times cited" of "Double Dipping". Double-dipping is initiated at the beginning of the research when funds and research workers first come together and this is captured in the "Input" Component 1 of this Principal Component Analysis. In this paper the loadings of these input funding variables appear in Component 1, and are far less important in the Component 3 where the loading of Times Cited is dominant – and vice versa. These two Components are orthogonal (not correlated), which suggests that the difficulty of seeing double-dipping behaviour using only "Times Cited" as a response measure is because of this lack of correlation and the low contribution it makes to the overall variability in the data set.

References

Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: A network-analytic model. *American Sociological Review, 63*(6), 829-846.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A Principal Component Analysis of 39 Scientific Impact Measures. *Plos One, 4*(6). doi: 10.1371/journal.pone.0006022

Bornmann, L. (2015). Letter to the Editor: On the conceptualisation and theorisation of the impact caused by publications. *Scientometrics, 103*(3), 1145-1148. doi: 10.1007/s11192-015-1588-4

Chen, K. H., Tang, M. C., Wang, C. M., & Hsiang, J. (2015). Exploring alternative metrics of scholarly performance in the social sciences and humanities in Taiwan. *Scientometrics, 102*(1), 97-112. doi: 10.1007/s11192-014-1420-6

Cooper, B. A., & Weekes, A. J. (1983). *Data, Models and Statistical Analysis*. Deddington, Oxford: Philip Allan.

Davey Smith, G., Sterne, J. A., Fraser, A., Tynelius, P., Lawlor, D. A., & Rasmussen, F. (2009). *The association between BMI and mortality using offspring BMI as an indicator of own BMI: large intergenerational mortality study* (Vol. 339).

Dimitrov, J. D., Kaveri, S. V., & Bayry, J. (2010). Metrics: journal's impact factor skewed by a single paper. *Nature, 466*(7303), 179-179. doi: 10.1038/466179b

Faba-Perez, C., Guerrero-Bote, V. P., & De Moya-Anegon, F. (2003). Data mining in a closed Web environment. *Scientometrics, 58*(3), 623-640. doi: 10.1023/b:scie.0000006884.08036.73

Faba-Perez, C., Guerrero-Bote, V. P., & De Moya-Anegon, F. (2004). Methods for analysing web citations: A study of web-coupling in a closed environment. *Libri, 54*(1), 43-53. doi: 10.1515/libr.2004.43

Fisher, L. D., & van Belle, G. (1993). *"Biostatistics, a Methodology for the Health Sciences*. New York: Wiley-Interscience.

Gingras, Y. (2014). Criteria for Evaluating Indicators. In B. Cronin, and Sugimoto, C. (Ed.), *Beyond Bibliometrics: Harnessing Multi-dimensional Indicators of Scholarly Impact*. Cambridge Massachusetts: MIT Press.

Hendrix, D. (2008). An analysis of bibliometric indicators, National Institutes of Health funding, and faculty size at Association of American Medical Colleges medical schools, 1997–2007. *Journal of the Medical Library Association : JMLA, 96*(4), 324-334. doi: 10.3163/1536-5050.96.4.007

Lewison, G., & Dawson, G. (1998). The effect of funding on the outputs of biomedical research. *Scientometrics, 41*(1-2), 17-27.

Manly, B. F. J. (1986). *Multivariate Statistical Methods – A Primer*: Chapman and Hall.

Oxford Dictionary of Modern Quotations. (2007). page 56.

Rigby, J., & Julian, K. (2014). On the horns of a dilemma: does more funding for research lead to more research or a waste of resources that calls for optimization of researcher portfolios? An analysis using funding acknowledgement data. *Scientometrics*, 1-9. doi: 10.1007/s11192-014-1259-x

Sample, I. (2013). Nobel winner declares boycott of top science journals *The Guardian*.

Setti, G. (2013). *Use and Misuse of Impact Factor and Other Bibliometric Indicators: What the IEEE Is Doing to Address This Issue?* Paper presented at the CDC 2013, Florence, Italy,.

The American Society for Cell Biology (ASCB). (2013). San Francisco Declaration.