Contents lists available at ScienceDirect

Journal of Informetrics

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

Citations to scientific articles: Its distribution and dependence on the article features

E.S. Vieira, J.A.N.F. Gomes*

REQUIMTE/Departamento de Química, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal

ARTICLE INFO

Article history:
Received 14 April 2009
Received in revised form 9 June 2009
Accepted 10 June 2009

Keywords:
Citations analyses
Field
Impact

ABSTRACT

The citation counts are increasingly used to assess the impact on the scientific community of publications produced by a researcher, an institution or a country. There are many institutions that use bibliometric indicators to steer research policy and for hiring or promotion decisions. Given the importance that counting citations has today, the aim of the work presented here is to show how citations are distributed within a scientific area and determine the dependence of the citation count on the article features. All articles referenced in the Web of Science in 2004 for Biology & Biochemistry, Chemistry, Mathematics and Physics were considered.

We show that the distribution of citations is well represented by a double exponential-Poisson law. There is a dependence of the mean citation rate on the number of co-authors, the number of addresses and the number of references, although this dependence is a little far from the linear behaviour. For the relation between the mean impact and the number of pages the dependence obtained was very low. For Biology & Biochemistry and Chemistry we found a linear behaviour between the mean citation per article and impact factor and for Mathematics and Physics the results obtained are near to the linear behaviour.

1. Introduction

Citation analysis has been used to assess the importance of a scientific publication as the number of times that it has been cited by other authors may be a good proxy for its overall scientific impact on the global community. Assuming that citation counts are related to scientific quality, they are currently used by decision makers to evaluate the academic performance of researchers (Ventura & Mombrú, 2006) or departments and research institutions (Adam, 2002; Bayer & Folger, 1966) when they have to take decisions related to funding and promotion as well to evaluate the scientific results achieved by countries (May, 1997; King, 2004).

Databases such as Thomson Reuters (Web of Science, WoS), Scopus, Google Scholar, Chemical Abstracts, American Psychological Association (PsycINFO) and American Mathematical Association (MathSciNet), allow users to search the citations obtained by each single paper. For many years the WoS was the quasi sole source of citation indexing, but this possibility is now common to the other databases (Vieira & Gomes, 2009). This may contribute to extend the applicability of bibliometric analysis to fields less well represented in the more comprehensive databases WoS and Scopus.

* Corresponding author. Tel.: +351220204507.
E-mail address: jfgomes@fc.up.pt (J.A.N.F. Gomes).

1751-1577/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2009.06.002

Several attempts are recorded in the literature to describe the citation distribution. A power law type distribution has been suggested (Naranan, 1971; Seglen, 1992), but no satisfactory theoretical model for the citation distribution exists so far. In the late 1980s, the Budapest group (Glänzel, Schubert, Telcs, Braun) used the Waring distribution to describe the bibliometric productivity (Braun, Glanzel & Schubert 1990). Sichel (1992) suggested the three-parameter Generalized Inverse Gaussian Poisson Distribution as a satisfactory mathematical model for the observed distribution of the number of references in a document and for other bibliometric properties. More recently, van Raan (2001) proposed a two-step competition process that leads to a modified Bessel-function distribution. Radicchi, Fortunato, and Castellano (2008) uses the lognormal distribution function to fit data on 14 among more than 200 subject categories used in WoS, mostly for the year 1999 with good results if the low citation points and the high tail is left out. Bornmann and Daniel (2009) studied the distribution citation rates, Radicchi et al.'s relative indicator (c_f), and z-scores (which have been used for many years in psychological testing for normalization of test scores) for the manuscripts accepted or rejected (but published elsewhere) by Angewandte Chemie International Edition. The finding indicates that z-scores are better suited than c_f values for a cross-discipline comparison of citation impact of publications.

The citations achieved by a paper may be affected by many factors, beyond its actual scientific content. The citation culture of the discipline does certainly have a macro-effect but other factors are likely to be important. Bornmann and Daniel (2008a) made a review of the studies on the citing behaviour of scientists in order to determine which factors lead a scientist to cite other papers. Authors compete to publish in the "best journals" believing that this will improve their impact in fellow researchers but the recognition of the author's name or institutional address may be relevant as well. Citation practice varies considerably among fields and thus the citation counts obtained can be much different. Radicchi et al. (2008) estimated the mean citation rate for some of the subject categories used in Journal Citation Report to show how much it varies across disciplines. Other variables that can be considered are the intrinsic characteristics of the publication, such as the number of co-authors, number of addresses, number of pages, number of references and impact factor of the journal. Glanzel and Thijs (2004) showed that the mean citation rate of papers in Biomedical Research, Chemistry and Mathematics grows with the number of co-authors, an effect particularly important for foreign citations. Leimu and Koricheva (2005) considered papers in 53 ecological journals to conclude that the mean citation rate is higher for 4 co-authors than for 3, 2 or one. Figg et al. (2006) looked at a set of journals with high impact factor to conclude that those researchers who are open to collaborations (papers with various authors) produce superior outputs that result in a higher impact. However, Herbertz (1995) finds no correlation between the number of addresses and the citation average of the publications from a number of research institutes in molecular biology. The influence of the length of a paper in the citation count was analyzed by Peters and van Raan (1994) considering a set of 226 papers written by 18 internationally recognized scientists in Chemical Engineering. The results showed that there is a relation between article length and impact at the shorter side of the length distribution but no significant relation was found for longer papers. Bornmann and Daniel (2007) studied the articles published by a set of 96 applicants for a research fellowship of an international foundation for the promotion of basic research in biomedicine. The results demonstrate that the number of pages can increase the effect on total citation counts. Peters and van Raan (1994) demonstrated also that the number of references correlates with the impact. The relation between the impact factor of the journal and the mean citation of the articles it publishes was studied in detail by Seglen (1994) for publication lists provided by 16 senior scientists of a major Norwegian biomedical research institute. For these lists he found that there is a poor correlation between article citedness and journal impact for the whole article population, but grouping the articles into defined journal impact cohorts seemed to improve the correlation. Other studies suggest that there may be a relationship between mean citation rate and the impact factor. Boyack and Klavans (2005) have selected a set of data combining SCIE/SSCI (Web of Science) for the years 2002 and 2003 and showed that there is a tendency for citations to increase as the journal impact factor increases. It has been pointed out (Bornman & Daniel, 2008b) that the journal impact factor should not be used in a study of this type as a possible explanation of the actual citedness of papers as it is calculated as an average of the actual citations obtained by papers in that journal. Our justification for using this parameter is that the final goal of this research is to seek a predicting mechanism for the future impact of a current paper and, from this point of view; the "quality" of the paper is certainly very relevant and the impact factor of the journal where it is published is perhaps the best proxy of the "quality" of the paper.

This study presents a very simple model for the citation distribution where citations are assumed to be randomly distributed among articles organized in classes of exponentially decreasing impact. This is of course a crude approximation as we believe that papers may be somehow differentiated by their intrinsic "quality" but we all know that there is no way of assessing *a priori* this "quality". The study considers the full set of articles referenced in 2004 in the WoS for the scientific areas of Mathematics, Physics, Chemistry and Biology & Biochemistry. These disciplines were selected for their large number of documents and different publication cultures. The aim of this work is to present a systematic study of the correlations that can be found between the impact of the articles and each of the variables (co-authors, addresses, pages, references and impact factor) when using a large volume of data in different fields. The effect of each variable was studied separately for each scientific field as the influence of these variables in the impact may depend on the field as earlier studies appear to suggest.

In the next section below, the methodology is described in detail. The following section contains the analysis of the distribution of citations in each field and the correlation between the mean citation rate and each of the article features. The final section summarizes the major conclusions that can be drawn from the study.

2. Methods

The study is based on the analysis of 226,166 articles published, in 2004, in journals indexed in the Web of Science (WoS) and classified according to the Essential Science Indicators (ESI) in the fields of Biology & Biochemistry, Chemistry, Mathematics and Physics (Thomson Reuters, 2009a). The ESI is a compilation of statistical information related with publications, citations and cites per paper for journals, scientists, institutions and countries referring to 10 years of Thomson Reuters data. The number of journals belonging to each field was 455 (44,248 articles) for Biology & Biochemistry, 574 (97,177 articles) for Chemistry, 387 (20,127 articles) for Mathematics and 338 (64,614 articles) for Physics. All information about these articles was collected from the WoS to build a database where we have information about co-authors, institutional addresses, number of pages, number of references, journals and number of citations (from year of publication to present)[1] for each document classified as an article. To study the distribution of citations we counted the citation of each individual article and then aggregated the articles with the same number of citations from zero to the maximum number of citations for that particular set of articles. To determine how the variables that characterize one article can influence the citations it achieves, the relation between the mean citation rate and the number of co-authors, number of addresses, number of pages, number of references and the journal impact factor were studied. For each field the articles with one, two, three, etc., co-authors were selected and their citations counted. (Articles where the author's field was identified as anonymous were eliminated.) The same methodology was applied for the other variables. The study of the relationship between the average citation per paper and the impact factor was made extracting the impact factor of 2006 Journal Citation Reports (JCR) for each journal belonging to the field analyzed.

3. Results and discussion

3.1. The distribution of citations

Fig. 1 shows the frequency of occurrence of articles with zero, one, two, three citations and so on for a 5-year citation window.

The distributions for Physics and Chemistry are similar. In Mathematics, citations are much less common, while for Biology & Biochemistry much higher citation counts occur. The parameters of these distributions are summarized in Table 1 below.

The distributions are very skewed but the medians follow closely the averages, the ratio varying from 0.52 to 0.66. The origin of these differences must be associated with the publication cultures of the disciplines. Basically Mathematicians give far less references in their articles so that the citation counts must be lower. Within each discipline (or sub-discipline) the reasons for a paper to receive a larger or smaller number of citations are difficult to establish in detail. If it is assumed that a citation means that the cited document influenced the citing author, the citation count would be a measure of the influence that document had in the scientific community. However, other reasons may lead to a citation so that the causal argument becomes much more difficult. It may be expected that dealing with such large numbers of documents, the study of the distribution may help understanding the motivations for citation.

3.1.1. 1st model: random citation

Assuming that an author decides to choose randomly his references, any published article would have the same (very small) probability of being cited. The number of citations obtained by each article would follow a Poisson distribution with the mean calculated as the ratio of the aggregate of all citations (of the articles published in 2004 by any publications referenced in the WoS in 2004–2008) to the number of articles in 2004. The Poisson distribution has a variance equal to the mean and the probability density function is

$$P(k;\mu)=\frac{\mu^k}{k!}e^{-\mu}$$

In Table 1 we can see that the results obtained for the variance are much larger than those of the mean, suggesting that the hypothesis that all articles have the same expected impact is wrong as we might have guessed. This result begs the question about the distribution of the expected impacts of papers in a certain discipline.

3.1.2. 2nd model: the expected impact of papers decreases exponentially

We can now assume that the articles of a given discipline can be organized in sets with the same likelihood of being cited or expected citedness. For the sake of the argument, we can assume that each author writes very few good papers and a large number of low quality papers, quality meaning here simply its expected impact. Let us assume that this quality is measured by the average citation count of the papers in this class. For each of these classes, the Poisson distribution would apply to describe the distribution of the number of citations that these equally good papers would get. The observed distribution would be the convolution of the Poisson distribution density function with that of the distribution of the average, its unique

[1] The data were extracted from the WoS in 9 to 13 January of 2009 so that the citation period for papers published in 2004 is very close to the 5-year window.

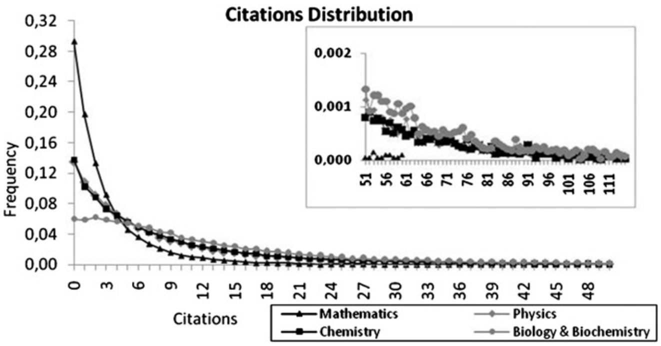

Fig. 1. The distribution of citations of the articles in the WoS published in 2004 for Biology & Biochemistry, Chemistry, Mathematics and Physics. Most Physics points are not seen as they lie behind the Chemistry line.

Table 1
Statistical parameters of the citation distribution of the 2004 articles in the WoS.

Field	No. articles	Average citations	Median	Variance	Standard Deviation
Biology & Biochemistry	44,248	13.59	9	358.34	18.93
Chemistry	97,177	9.55	5	284.93	16.88
Mathematics	20,127	3.22	2	38.44	6.20
Physics	64,614	9.72	5	269.62	16.42

parameter. This average number of citations will be called here the expected citedness of papers in one particular class. Let us assume that the distribution of this expected citedness, μ, has an exponential form,

$$F(\mu; E) = \frac{1}{E} e^{-(\mu/E)}$$

where the parameter E is the mean of the random variable μ that is, the mean of the expected citedness of all classes of articles assumed to form a continuous set. The probability of occurrence of a paper with k citations may be calculated by direct integration of Poisson distribution and Exponential distribution

$$P(k) = \int_0^{\alpha} P(k; \mu) F(E; \mu)\, d\mu$$

that may easily be shown to be given by

$$P(\mu) = \frac{1}{E+1} \left(\frac{E}{E+1} \right)^k$$

This distribution has a mean E and a variance equal to $E(E+1)$. This may be called exponential-Poisson mixture or exponential-Poisson distribution to make clear the citation model we are using. In fact, this distribution of the probability of occurrence of a paper with k citations is just the well known geometric distribution. We used this distribution for the four disciplines considered as shown in Fig. 2, where the parameter E was not fitted but chosen to be equal the observed mean number of citations.

The quality of the fit is surprisingly good, especially considering that the exponential-Poisson distribution has a unique parameter and that we used the average value of the citations per article without any adjustment. This suggests a degree of validity of the hypothesis that articles may be ranked by a property somehow linked to their quality, the expected impact as measured by the expected number of citations or citedness; this number was assumed to be distributed as an exponential decay as shown in Fig. 3 below, for the disciplines we considered and using the average number of citations observed.

As expected Physics and Chemistry are very similar but Mathematics, on the one side, and Biology & Biochemistry, on the other, behave differently. In Mathematics, the frequency of occurrence of articles with low expected citedness is very high while in Biology & Biochemistry higher expected citedness is more common. The top 10% of the articles in terms of their expected citedness lie above 31 in Biology & Biochemistry, above 22 in Physics or Chemistry and above 7 in Mathematics. These striking differences in the position of the expected citedness deciles are shown in Table 2.

It is surprising how this simple assumption leads to citation distributions so close to those observed noting that the particular culture of each discipline is represented by nothing more than the average number of citations per article. If we want to improve the quality of these fits, a second parameter should be introduced in the distribution function.

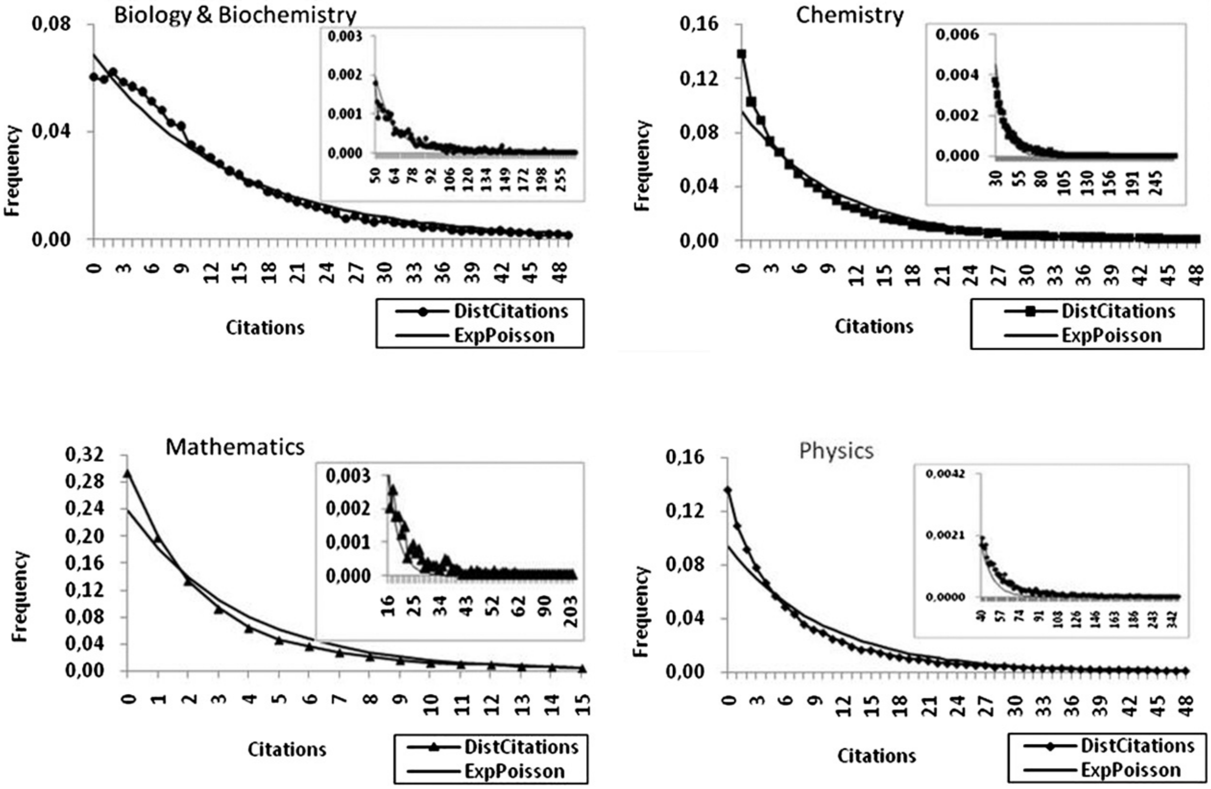

Fig. 2. The exponential-Poisson distribution adjusted to the empirical data for Biology & Biochemistry, Chemistry, Mathematics, and Physics.

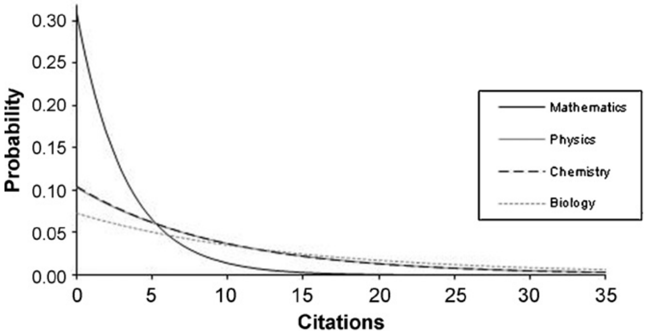

Fig. 3. Exponential distributions of the expected citedness of articles for Biology & Biochemistry, Chemistry, Mathematics and Physics. Physics points lie behind the Chemistry line.

Table 2
Decile positions and average citedness for the four disciplines considered.

Decile	Decile positions				Citedness average in each decile			
	Biology & Biochemistry	Chemistry	Mathematics	Physics	Biology & Biochemistry	Chemistry	Mathematics	Physics
1 st	1.43	1.01	0.34	1.02	0.703	0.494	0.167	0.503
2nd	3.03	2.13	0.72	2.17	2.216	1.557	0.525	1.585
3rd	4.85	3.41	1.15	3.47	3.919	2.754	0.928	2.803
4th	6.94	4.88	1.64	4.96	5.867	4.123	1.390	4.196
5th	9.42	6.62	2.23	6.74	8.142	5.721	1.929	5.823
6th	12.45	8.75	2.95	8.90	10.877	7.644	2.577	7.780
7th	16.36	11.50	3.87	11.70	14.311	10.056	3.390	10.236
8th	21.87	15.37	5.18	15.64	18.927	13.301	4.483	13.538
9th	31.29	21.99	7.41	22.38	26.037	18.297	6.167	18.623
10th	∞	∞	∞	∞	44.873	31.533	10.629	32.096

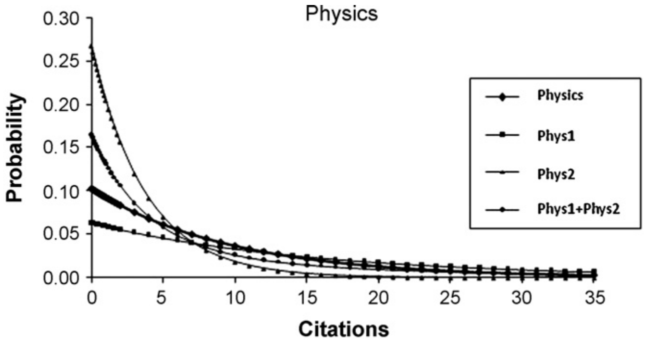

Fig. 4. Exponential distributions of the expected citedness of articles for Physics: comparison of the single exponential with the double exponential distribution.

3.1.3. 3rd model: double exponential-Poisson distribution

In an attempt to improve the fitting of our model to the citation data, we try a more flexible distribution for the expected citedness obtained by the average of two exponentials. The effect is shown in Fig. 4 for the Physics case. The effect of this two-parameter distribution may be understood by looking at the new deciles. The lowest 10% expected citation is reduced from citedness 1.02–0.55; at 50%, the citedness decreases from 6.74 to 3.90; at the other end, the 90% highest expected citation decile increases from 22.4 to 24.0. Of course, the simple exponential-Poisson (or geometric) distribution and the new double exponential lead to the same average citation impact of the discipline considered. The net result is that of increasing the frequency of low cited articles while decreasing the predicted frequency in the higher citation region.

Using this new distribution with the conditions that it should reproduce the average and the standard deviation of the data, we obtain the fits shown in Fig. 5 below. The quality of the fit improves markedly both in the short and in the medium range and also in the long tail. An alternative procedure consists of doing a least squares fit of the double exponential-Poisson expression to the empirical data. This improves slightly the adjustment without any major qualitative difference.

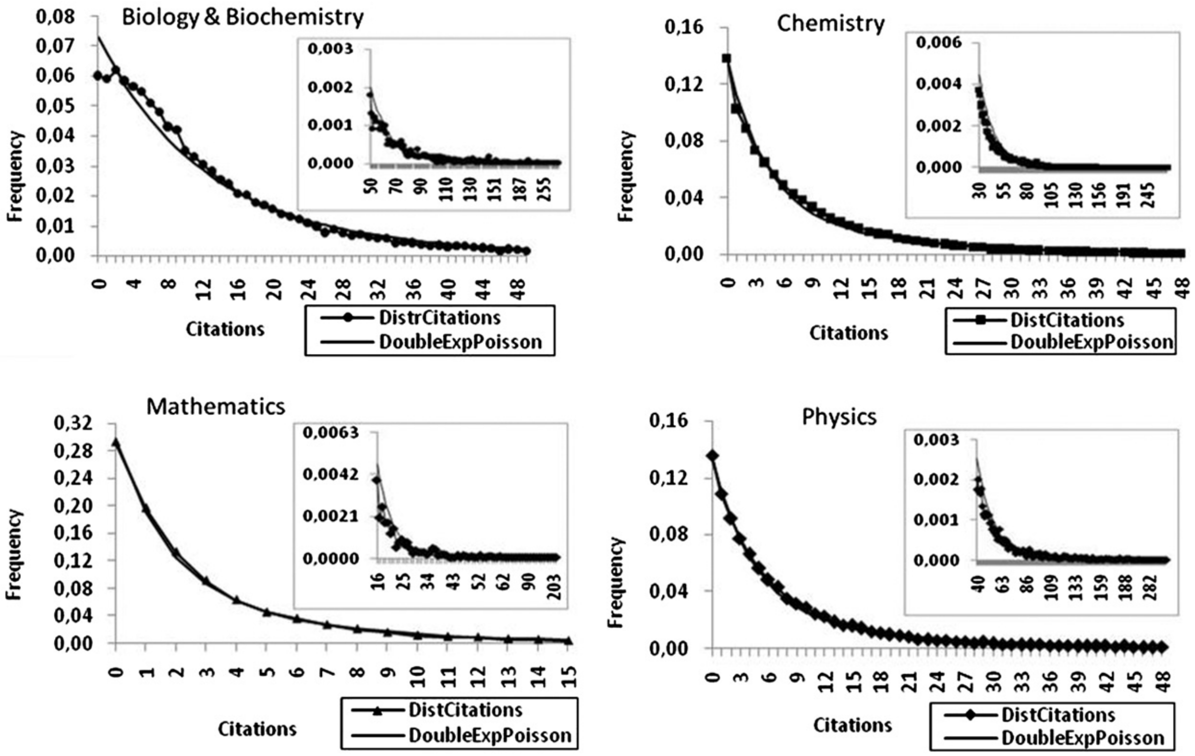

Fig. 5. The double exponential-Poisson distribution adjusted to the empirical data for Biology & Biochemistry, Mathematics, Physics and Chemistry.

Table 3
The Kolmogorov–Smirnov test.

Field	D_{obs}		Critical values for the K–S, D_{tab}	
	Exponential Poisson	Double Exp. Poisson	Significance level 0.01	Significance level 0.05
Biology & Biochemistry	0.034	0.024	0.115	0.096
Chemistry	0.075	0.019	0.107	0.089
Mathematics	0.073	0.015	0.181	0.151
Physics	0.092	0.022	0.110	0.092

To assess the quality of the fit we used the Kolmogorov–Smirnov (K–S) test. The K–S is used to test the hypothesis that a given data set could have been drawn from a given distribution. The statistic of the test is given by the equation:

$$D_{obs} = Max|F(x) - S(x)|$$

where $F(x)$ and $S(x)$ are the theoretical and observed distribution functions, respectively.

Table 3 confirms that the proposed distributions cannot be ruled out with the double exponential-Poisson distribution improving the description as justified above.

It has been suggested that a negative binomial distribution (or Gamma-Poisson) would fit well the citation data but all the tests were done with a fairly limited number of documents (Bornmann & Daniel, 2006; Glanzel & Schubert, 1991). Of course, this negative binomial distribution corresponds to the geometric distribution (or single exponential-Poisson distribution) when one of its two parameters is fixed. These simple models suggest that the citation success of documents depends on properties that can be reduced to a single number, the expected citedness, this having a distribution that can be represented by the simple average of two exponentials (the double exponential in our study) or a Gamma distribution. This finding does not void the possibility that particular features of the document may influence its future citations. The journal where it is published is expected to have a major influence as all authors scramble to have their publications in the best journals for their particular subjects. Other features such as the size of the paper or its number of pages, the number of references, the number of co-authors or the number of addresses may conceivably influence the number of citations it will get. These properties may be measured in an obvious way if we disregard the very different size of text and figures that may fit into a page of a journal depending on its format. The most frequently used parameter to measure the quality of a journal is the WoS impact factor that is calculated by dividing the number of citations in the Journal Citation Report year by the total number of articles published in the 2 previous years (Thomson Reuters, 2009b). In this paper, we count citations of 2004 publications over a 5-year period. The average of this count does not coincide with the impact factor but will be related with it in a less obvious way. The dependence of the average says very little about the citation behaviour of single documents as the variance is very large and the distributions are skewed, justifying further study.

3.2. How the citation count depends on the features of the article

We consider now the possible dependence of the citation count of an article on some of its features, namely the number of co-authors, the number of addresses, the number of pages, the number of references and the impact factor of the journal where it was published. For each of the four scientific fields considered, we plot the mean citation rate against the magnitude of the parameter. The dependence of the mean citation rate on the number of co-authors for each field is shown in Fig. 6, where the estimated standard deviation of the mean is shown as an error bar.

The results suggest a dependence between the mean impact and the number of co-authors. The correlation coefficients, R^2, are low as a result of the peculiar behaviour of the one-author documents and also of the scatter of the points associated with larger numbers of co-authors for which the number of documents is low and the error of the mean high. The maxim number of co-authors found was 27 in Biology & Biochemistry and in Chemistry, 12 in Mathematics and 592 in Physics. For numbers of co-authors higher than those showed in the figure, the number of documents becomes much lower and the errors associated are more pronounced. The average number of co-authors per document is the highest for Biology & Biochemistry (4.84) and almost the same for Chemistry and Physics (3.98 and 3.89, respectively). For Mathematics the average number found was 1.84 co-authors per document. For Biology & Biochemistry, Chemistry and Physics we can see that the mean citation rate for articles with one author is lower than what might be expected by extrapolation of the values found for two or more authors. For Mathematics this value falls in the straight line drawn for higher numbers.

Fig. 7 shows the mean impact of the articles against the number of addresses for Biology & Biochemistry, Chemistry, Mathematics and Physics, with the estimated standard deviation of the mean shown as an error bar.

The dependence of the mean citation impact on the number of addresses is clear and the smaller correlation coefficient for Physics is due to the scatter of less common larger number of addresses. The average number of addresses per document is almost the same in all fields (2.25, 1.75, 1.62 and 2.11 for Biology & Biochemistry, Chemistry, Mathematics and Physics, respectively). The maximum number of addresses was found for Physics at 114, while for the other fields the maximum obtained were much lower (22, 10 and 7 addresses for Biology & Biochemistry, Chemistry and Mathematics, respectively). Contrary to the previous case, the points corresponding to single address articles are very close to the regression line but we can still argue that they show a special behaviour for Chemistry and Physics.

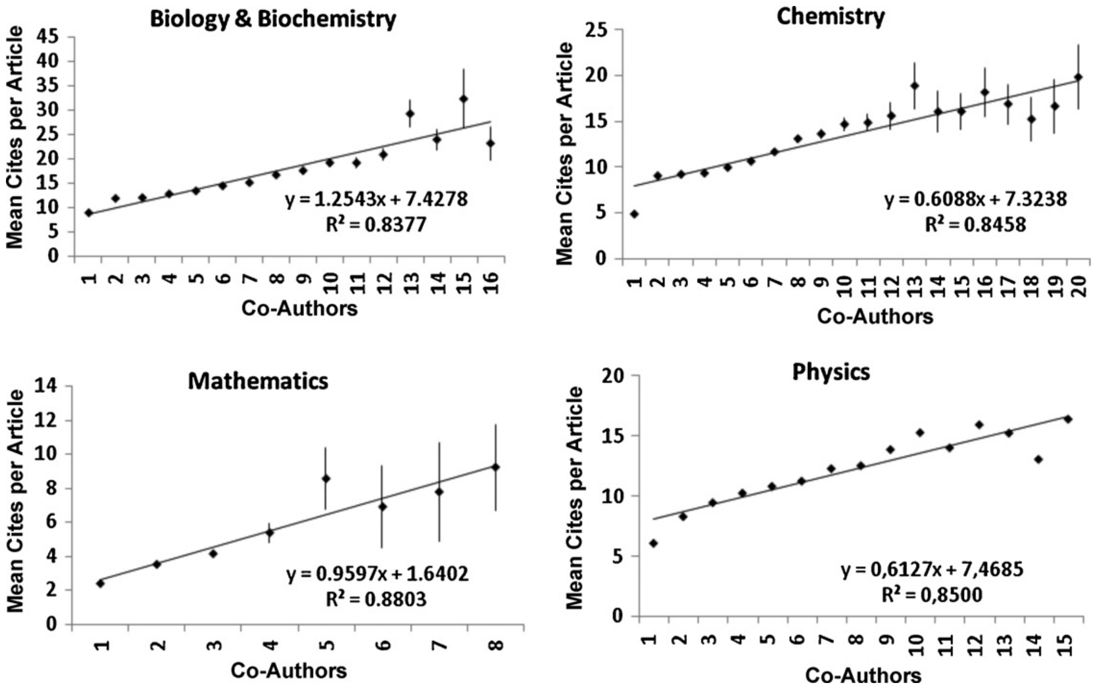

Fig. 6. Impact obtained by articles published in 2004 against number of co-authors in Biology & Biochemistry, Chemistry, Mathematics and Physics. Each point represents the mean citation count of all articles with that number of co-authors.

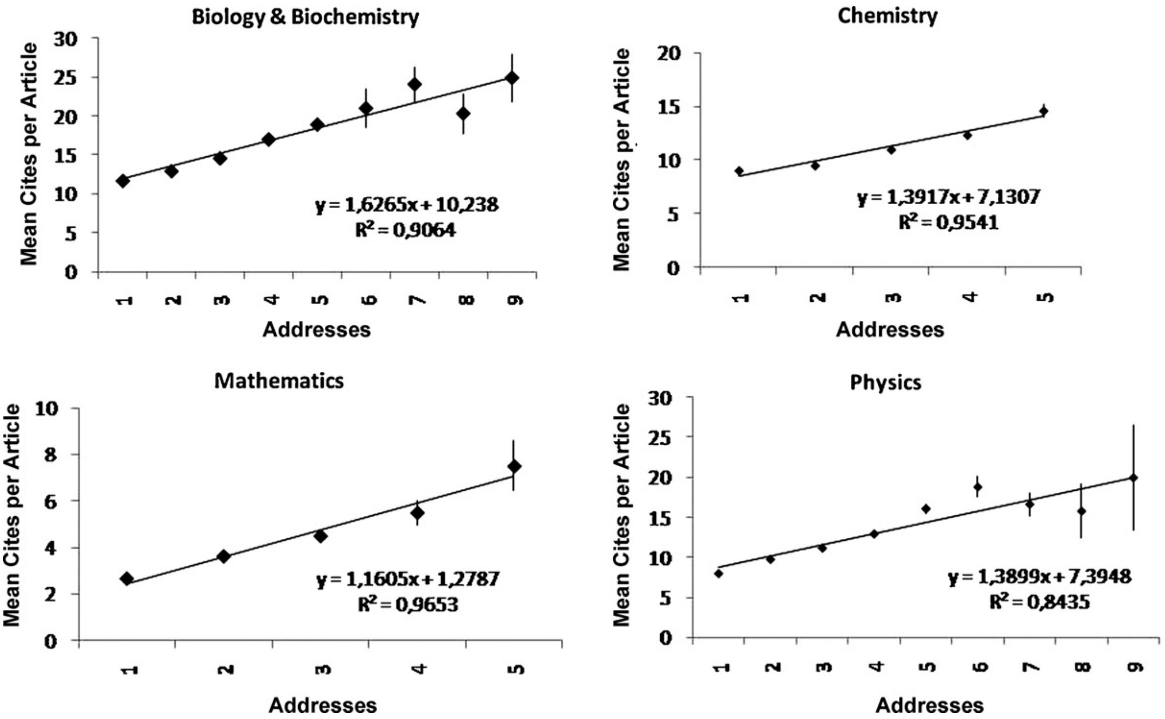

Fig. 7. Impact of the articles published in 2004 plotted against the number of addresses in Biology & Biochemistry, Chemistry, Mathematics and Physics. Each point represents the mean citation count of all articles with that number of addresses.

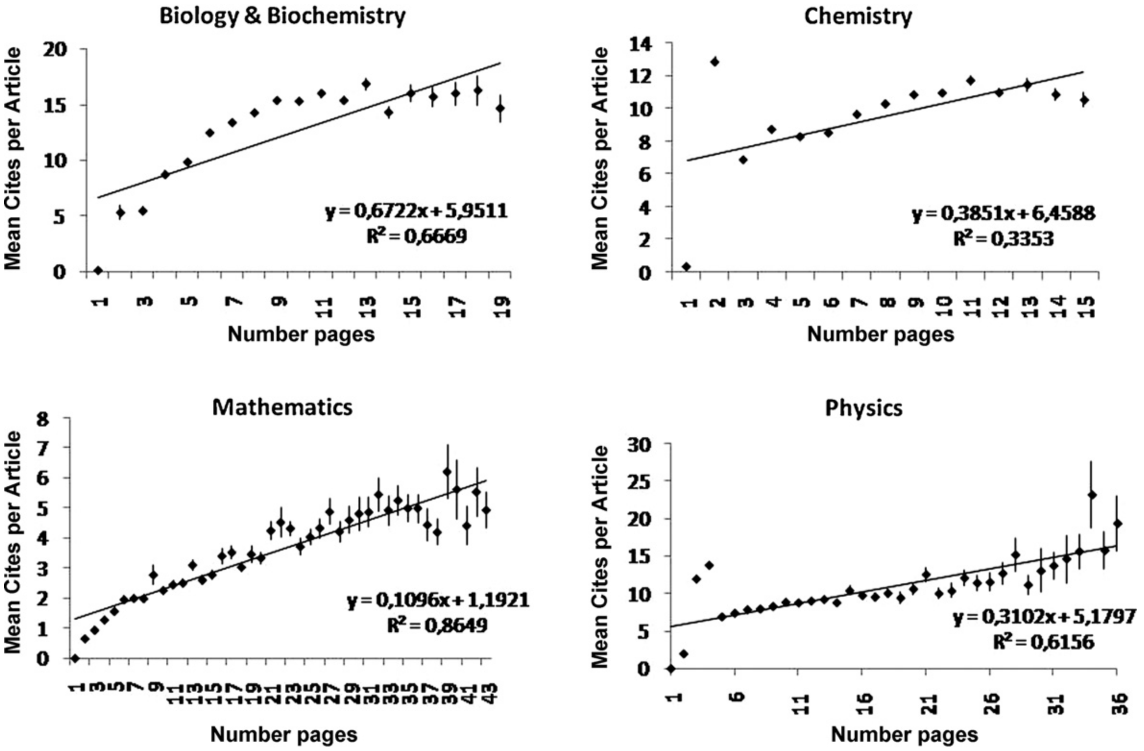

Fig. 8. Impact of the articles published in 2004 against number of pages in Biology & Biochemistry, Chemistry, Mathematics and Physics. Each point represents the mean citation count of all articles with that number of pages.

Fig. 8 shows the plot of the mean impact of the articles against the number of pages for Biology & Biochemistry, Chemistry, Mathematics and Physics, with the estimated standard deviation of the mean shown as an error bar.

The dependence of the citation impact on the number pages is far more complex. This is likely to be explained by the different nature of the articles indexed in WoS for these disciplines. Documents with one two or three pages are classified as "articles" in WoS that would not normally be considered primary research papers and the citations they elicit fall far from the pattern of other articles. The average number of pages per document is 17.82, 9.72, 8.68 and 6.89 for Mathematics, Physics, Biology & Biochemistry and Chemistry, respectively. The maximum number of pages found was 512, 333, 240 and 198 for Mathematics, Chemistry, Physics and Biology & Biochemistry, respectively. The number of pages is perhaps a good indicator of other features that define different classes of documents that are classified by Thomson Reuters as articles. In all fields, single page articles have a very small impact, far below that expected from the regression line. Looking in detail to the documents classified in this group, we find that this corresponds to short news items published in several journals and classified by WoS as articles. In other cases, it includes papers delivered at conferences and published in Physics Today, Physics World, Chemical Week and Chemical World among others. In all cases, the number of single page articles is small but this does not show in the error bar as they have a consistently small impact. The mean citation rate obtained for articles with two pages for Chemistry and three or four pages for Physics is relatively high. These are typically published in journals with high impact factor. For Chemistry, the majority of these papers were published in journals like Angewandte Chemie-International Edition, Chemical Communications and Journal of the American Chemical Society classified by the journal as communications while in the WoS they are considered articles. For Physics, these papers were published in journals like Applied Physics Letters, Optics Letters, IEEE Photonics Technology Letters or Journal of Applied Physics where the average number of pages per document is between three and four.

Fig. 9 shows the plot of the mean impact of the articles against the number of references for Biology & Biochemistry, Chemistry, Mathematics and Physics, with the estimated standard deviation of the mean shown as an error bar.

The dependence is very clear but a little off the linear behaviour and this is translated in relatively low correlation coefficients. The average number of references per document is the highest for Biology & Biochemistry (36.76). For Chemistry, Physics and Mathematics the average found was 28.64, 24.11 and 18.25, respectively. For Biology & Biochemistry we can observe that the set of articles with 2 references have a mean impact far above what would be expected. This is due to a single article with 207 citations that influences the mean of just 37 articles with 2 references. This is a true outlier contributing to a large error bar. Articles with just a few references have a very small average number of citations, far below that predicted by the regression line in all areas but Mathematics where they come close to the prediction.

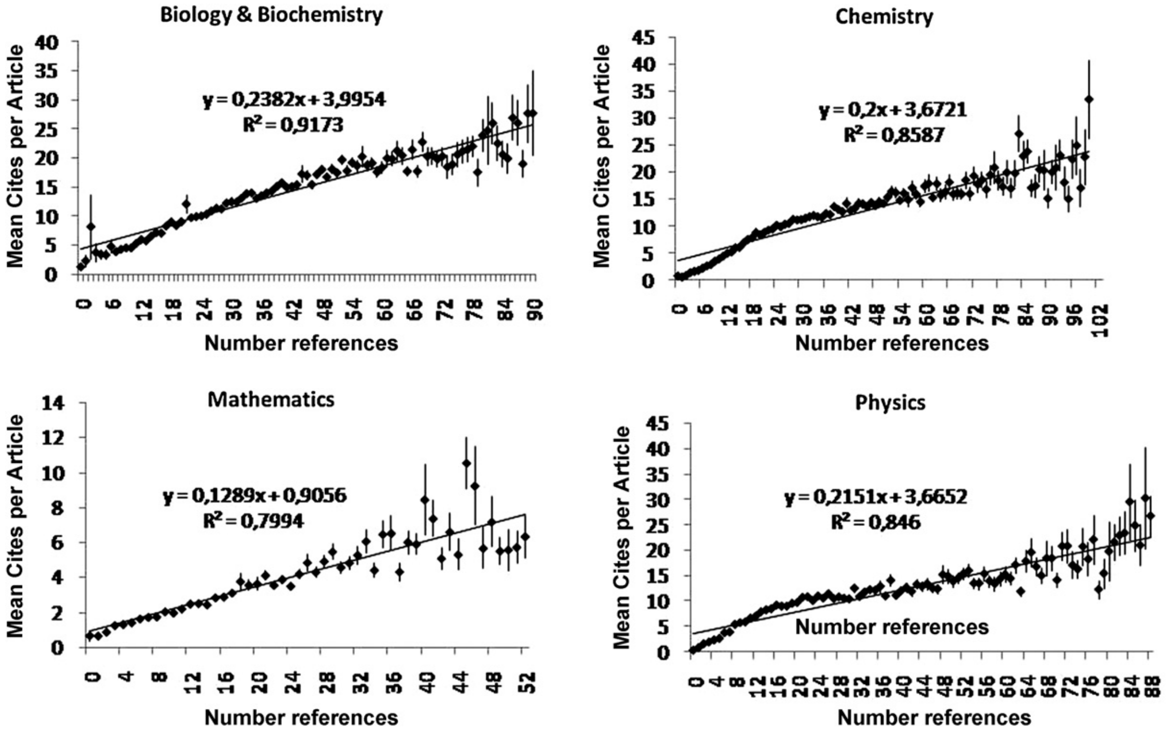

Fig. 9. Impact of the articles published in 2004 against the number of references in Biology & Biochemistry, Chemistry, Mathematics and Physics. Each point represents the mean citation count of all articles with that number of references.

For all fields, the increase in the standard deviation as the number of references grows is due to the smaller number of articles.

Fig. 10 shows the variation of the mean citation rate in the 5-year window considered with the impact factor of the journal. Error-bars are not shown as the standard deviation of the mean cites per article are too small to be seen in the plots.

The impact factor of the journals in this study varies between 0.048 and 24.007 for Biology & Biochemistry, between 0.051 and 10.232 for Chemistry, between 0.090 and 2.426 for Mathematics and between 0.121 and 7.072 for Physics. The mean impact factor is for Biology & Biochemistry 3.585, for Chemistry 2.435, for Mathematics 0.663 and for Physics 2.488. It should be noted that this mean is obtained as the ratio of the number of citations in 2004–2008 of the articles dated 2004 while Thomson Reuters considers all citations in 2006 divided by the number of articles and reviews published in 2004 and in 2005. The number of articles that are published in journals that are not in 2006 JCR is 3.2%, 0.84%, 0.35% and 2.7% of the total in Biology & Biochemistry, Chemistry, Mathematics and Physics, respectively. In Fig. 10 above, only points corresponding to impact factors with 50 or more articles (in one or more journals) are shown. It should be noticed that the lines in Fig. 10 might be forced to go through the origin but that would not change the conclusions as the slope suffer minor changes and the coefficient R^2 decreases by less than 0.01. The correlation coefficients go from 0.998 in Biology & Biochemistry to 0.93 in Mathematics. Considering that we are plotting two different measures of the same property, we would expect a higher correlation. However, using different time spans, introduces the effect of the time lag of the citations that is likely to vary with the impact factor of the journal. A similar scatter may be found in the plot of the WoS 5-year impact factor against the traditional 2-year impact factor.

In order to assess the inter-correlation between each of the five independent variables considered in this study, we determined the Spearman coefficients. For each field studied the highest values of the Spearman coefficient occur between the number of authors and the number of addresses [0.48–0.70] and between the number of pages and the number of references [0.50–0.57]. However, these values are not very high and suggest a moderate inter-correlation following Finney (1980). For the inter-correlation among the other variables, the values obtained for the Spearman coefficients suggest a very weak or weak correlation (Finney, 1980).

The values obtained for the Spearman coefficients suggest that the number of addresses and the number of pages should not be considered in future studies as independent variables. The number of references is to be preferred to the number of pages as this is too dependent on the printing format of the journal. As to the number of addresses, this may be considered to be dependent on the number of authors and not the other way around, this suggesting the more appropriate choice of independent variable between the two. This suggests that only one of these factors need to be considered in a study using citation data.

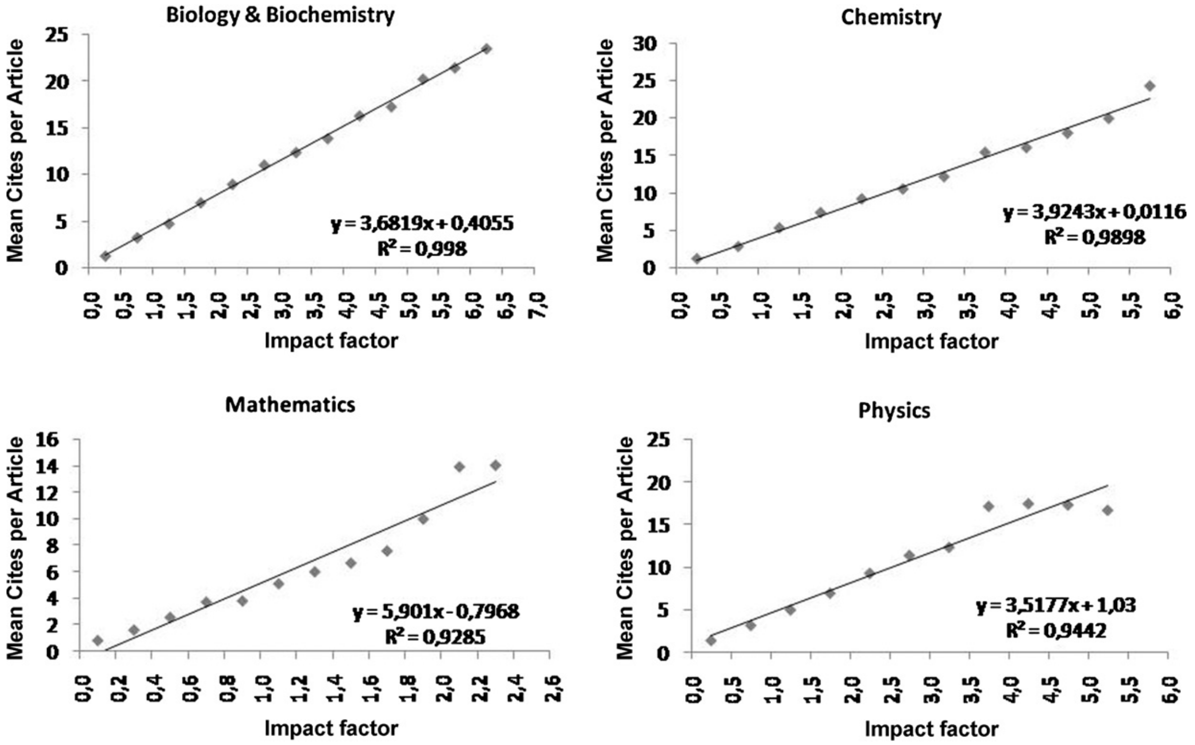

Fig. 10. Impact of the articles published in 2004 against impact factor of the journals in Biology & Biochemistry, Chemistry, Mathematics and Physics. Each point represents the mean citation count of all articles published in journals with impact factor within intervals of 0.5 or 0.2 for Mathematics.

4. Summary and conclusions

This paper reports on the distribution of citations of articles and how they depend on some of their own features. We considered the whole set of more than 220,000 articles published in 2004, in journals indexed in the WoS and belonging to the fields of Biology & Biochemistry, Chemistry, Mathematics and Physics. The number of co-authors, the number of institutional addresses, the number of pages, the number of references and the journal impact factor were considered as basic features that may have direct influence on the citation count. The distribution of citations for each of these variables and its effect upon the article citation count are summarized in Table 4.

The dependence found here between the citation counts and five properties of the documents is robust in the sense that we can predict that if this study were to be repeated for another year, for example on the citations of the 2003 publications in 2003–2008, the results would corroborate those reported here. On the other hand, the correlations found for the means of the citations of very large sets of documents do not allow direct conclusions for individual documents. This can be seen in the values of the standard deviation of the means plotted in Figs. 7–11 that are small but correspond to very large for the population, for the ensemble of documents used to form the mean.

The citation enhancement presented in Table 4 and Fig. 11 refers to the % increase in the mean citation count when we go from the mean to twice the mean for each independent variable. The impact factor is the variable with the largest effect. We see that the average citation per count of articles published in journals with an impact factor of twice the mean is close to the double of that of the articles published in journals with an impact factor coinciding with the mean, and this is true for the different scientific areas considered except for Mathematics where the increase is more than the double. For the impact factor, the very large value of the citation enhancement for Mathematics may be explained by the larger time lag of the citations in this discipline. Of course this is a direct result of Garfield's definition of journal impact factor as used in WoS. When we consider the effect of the number of references, the enhancement of the average citation per article varies from 58% in Physics to 72% in Mathematics. The increase of the number of pages from its average to twice this value produces a citation enhancement varying from 30% in Chemistry to 62% in Mathematics. For the number of addresses in the articles, the enhancement goes from 25% in Chemistry to 60% in Mathematics. Finally for the number of co-authors of the article, the enhancement varies from 24% in Physics to 52% in Mathematics. The lines in Fig. 11 represent each of the five variables and the intersection with *xx* and *yy* axis the percent enhancement for each field. Mathematics is seen to be the area with the largest effect for any of the five variables considered. The minimum effects are found for Physics and Chemistry. The impact factor is the variable with the highest enhancement, followed by the number of references and the number of pages. The enhancements for the number of co-authors

Table 4
Distribution of articles over some the article features and their effect of citation enhancement.

Variables	Biology & Biochemistry	Chemistry	Mathematics	Physics
Co-authors				
Average	4.84	3.98	1.84	3.89
Standard Deviation	2.65	4.40	0.90	38.98
R^2	0.837	0.845	0.880	0.850
Citation enhancement	45%	25%	52%	24%
Addresses				
Average	2.25	1.75	1.62	2.11
Standard Deviation	1.25	0.86	0.80	6.5
R^2	0.906	0.954	0.965	0.843
Citation enhancement	26%	25%	60%	28%
Pages				
Average	8.69	6.89	17.82	9.72
Standard Deviation	4.24	4.04	13.98	6.69
R^2	0.666	0.335	0.864	0.615
Citation enhancement	50%	30%	62%	37%
References				
Average	36.76	28.64	18.25	24.11
Standard Deviation	17.04	17.17	12.27	14.89
R^2	0.917	0.858	0.799	0.846
Citation enhancement	69%	60%	72%	58%
Impact factor				
Average	3.585	2.435	0.663	2.488
Standard Deviation	2.51	3.75	0.40	2.79
R^2	0.998	0.989	0.929	0.944
Citation enhancement	97%	99%	125%	89%

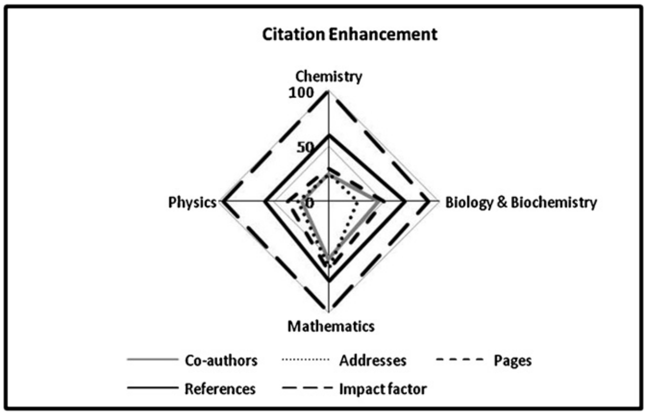

Fig. 11. Citation enhancement in percent of the initial value as a single variable goes from its average to twice the average.

and number of addresses are about the same in Chemistry, Mathematics and Physics but clearly different for Biology & Biochemistry.

The study presented here shows some dependence between the features of an article and the mean citation rate contrary to what other studies appeared to suggest (Herbertz, 1995; Seglen, 1994). Other conclusions are as follows.

(a) The average number of citations per article is the highest in Biology & Biochemistry, almost the same for Chemistry and Physics, and the lowest for Mathematics.
(b) The distribution of citations is well represented by a double exponential-Poisson distribution for all fields.
(c) The results showed a dependence of the mean citation count per article on some features of the articles, although this dependence may be far from the linear in some cases. This was also demonstrated by Bornmann and Daniel (2006) but they used the negative binomial regression model (NBRM).
(d) Biology & Biochemistry has the highest number of co-authors per article and Mathematics the lowest. For Biology & Biochemistry, Chemistry and Physics we can see that the mean citation rate for articles with one author is lower than what might be expected by extrapolation of the values found for two or more authors.
(e) The average number of addresses per document is almost the same in all fields.

(f) Single page articles have a very small impact in all fields. This is due to particular types of documents that WoS classifies as articles.
(g) The average number of references per document is the highest for Biology & Biochemistry and the lowest for Mathematics.

This paper reports on the dependence of the mean citation counts of very large sets of documents on certain features of the cited document. For these large sets, the standard deviations of the mean are small even when the standard deviations of the populations are high. This is well exemplified by the dependence on the impact factor of a journal: The standard deviation of the number of citations through 2008 of the articles in the Journal of the American Chemical Society in 2004 is as high as 31.5 for a mean value of 31.6, while this mean has an associated standard deviation of just 0.6. A word of caution may be appropriate as the conclusions drawn from the study of the means are not directly transferable to individual documents.

Acknowledgement

The financial support of Fundação para a Ciência e a Tecnologia (Lisbon) is acknowledged.

References

Adam, D. (2002). The counting house. *Nature*, *415*, 726–729.

Bayer, A. E., & Folger, J. (1966). Some correlates of a citation measure of productivity in science. *Sociology of Education*, *3*, 381–390.

Bornman, L., & Daniel, H. D. (2008). Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by Angewandte Chemie International EditionI, or rejected but Published Elswhere. *Journal of the American Society for Information Science and Technology*, *59*(11), 1841–1852.

Bornmann, L., & Daniel, H.-D. (2006). Selecting scientific excellence through committee peer review—a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, *68*(3), 427–440.

Bornmann, L., & Daniel, H. D. (2007). Multiple publications on a single research study: Does it pay? The influence of number of research articles on total citation counts in biomedicine. *Journal of the American Society for Information Science and Technology*, *58*(8), 1100–1107.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80.

Bornmann, L., & Daniel, H.-D. (2009). Universality of citation distributions—a validation of Radicchi et al.'s relative indicator $cf = c/c_0$ at the micro level using data from chemistry. *Journal of the American Society for Information Science and Technology*, *60*(6).

Boyack, K. W., & Klavans, R. (2005). Predicting the importance of current papers. In P. Ingwersen, & B. Larsen (Eds.), *Proc 10th int conf int soc scientometrics informetrics* (pp. 335–342). Stockholm: Karolinska University Press.

Braun, T., Glanzel, W., & Schubert, A. (1990). Publication productivity—from frequency-distributions to scientometrics indicators. *Journal of Information Science*, *16*(1), 37–44.

Figg, W. D., Dunn, L., Liewehr, D. J., Steinberg, S. M., Thurman, P. W., Barrett, J. C., et al. (2006). Scientific collaboration results in higher citations rates of published articles. *Pharmacotherapy*, *26*(6), 759–767.

Finney, D. J. (1980). *Statistics for biologists*. London: Chapman and Hall.

Glanzel, W., & Schubert, A. (1991). A characterization of scientometrics distributions based on harmonic means. In *European workshop on scientometric methods of research evaluation in the sciences, social sciences and technology*. Potsdam, Germany: Elsevier Science Bv.

Glanzel, W., & Thijs, B. (2004). Does co-authorship inflate the share of self-citations? *Scientometrics*, *61*(3), 395–404.

Herbertz, H. (1995). Does it pay to cooperate—a bibliometrics case-study in molecular biology. *Scientometrics*, *33*(1), 117–122.

King, D. A. (2004). The scientific impact of nations what different countries get for their research spending. *Nature*, *430*, 311–316.

Leimu, R., & Koricheva, J. (2005). What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution*, *20*(1), 28–32.

May, R. M. (1997). The scientific wealth of nations. *Science*, *275*, 793–796.

Naranan, S. (1971). Power law relations in science bibliography: A self-consistent interpretation. *Journal of Documentation*, *27*, 83–97.

Peters, H. P. F., & van Raan, A. F. J. (1994). On determinants of citations scores—a case-study in chemical engineering. *Journal of the American Society for Information Science*, *45*(1), 39–49.

Radicchi, F., Fortunato, 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.

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, *43*, 628–638.

Seglen, P. O. (1994). Causal relationship between article citedness and journal impact. *Journal of the American Society for Information Science*, *45*(1), 1–11.

Sichel, H. S. (1992). Note on a strongly unimodal bibliometrics size frequency distribution. *Journal of the American Society for Information Science*, *43*(4), 299–303.

Thomson Reuters. (2009a). *In cites, list of journals*. Retrieved January, 2008, from http://www.in-cites.com/journal-list/index.html.

Thomson Reuters. (2009b). *Journal citation reports*. Retrieved May, 2009, from http://admin-apps.isiknowledge.com/JCR/help/h_impfact.htm.

van Raan, A. F. J. (2001). Two-step competition process leads to quasi power law income distributions. Application to scientific publication and citation distributions. *Physica A*, *298*, 530–536.

Ventura, O., & Mombrú, A. W. (2006). Use of bibliometric information to assist research policy making. A comparison of publication and citation profiles of Full and Associate Professors at a School of Chemistry in Uruguay. *Scientometrics*, *69*(2), 287–313.

Vieira, E. S., & Gomes, J. A. N. F. (2009). A comparison of Scopus and Web of Science for a typical university. *Scientometrics*, on-line