

Big Macs and Eigenfactor Scores: Don't Let Correlation Coefficients Fool You

Jevin West¹ Theodore Bergstrom² Carl T. Bergstrom^{1,3}

November 10, 2009

¹Department of Biology, University of Washington, Seattle, WA

²Department of Economics, University of California, Santa Barbara, CA

³Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501

Keywords: EigenfactorTM Metrics, EigenfactorTM Score, Article InfluenceTM
Score, Impact Factor, Correlation Coefficient

Abstract

The Eigenfactor™ Metrics provide an alternative way of evaluating scholarly journals based on an iterative ranking procedure analogous to Google's PageRank algorithm. These metrics have recently been adopted by Thomson-Reuters and are listed alongside the Impact Factor in the Journal Citation Reports. But do these metrics differ sufficiently so as to be a useful addition to the bibliometric toolbox? Davis (2008) has argued otherwise, based on his finding of a 0.95 correlation coefficient between Eigenfactor score and total citations for a sample of journals in the field of medicine. This conclusion is mistaken; here we illustrate the basic statistical fallacy to which Davis succumbed. We provide a complete analysis of the 2006 Journal Citation Reports and demonstrate that there are important differences between the information provided by the Eigenfactor Metrics and that provided by Impact Factor and Total Citations.

1 Big Macs and Correlation Coefficients

One might be think that if the correlation coefficient between two variables is high, those variables convey the same information, and thus can be used interchangeably — but this line of reasoning is mistaken. A simple example helps to illustrate. In Table 1, we provide two statistics for each of 22 countries: the cost of a Big Mac in local currency, and the mean hourly wage in local currency. The Pearson product-moment correlation coefficient, ρ , between these two statistics is 0.99. Since ρ is nearly 1, one might conclude that we can use hourly wages to predict burger prices with high accuracy and one might question why anyone should waste his or her time collecting burger price information if the hourly wage rates are already known. But this line of reasoning is mistaken, as we can clearly see in the column labeled “Real Wage.” Even though the correlation coefficient between hourly wages and burger prices is $\rho = 0.99$, the critical quantity that determines a consumer’s purchasing power — the “real wage” defined as the ratio of hourly wages to burger prices — varies dramatically across countries. In Denmark, a worker making the mean hourly wage need only work for seven minutes to earn a Big Mac, whereas in China, a worker making the mean hourly wage must work for nearly two hours to afford a burger.

In our hamburger example, it is pretty clear what is going on. The denominations of currencies vary immensely and arbitrarily. It is indeed true that differences in real wages are small relative to differences in currency denominations. But it is not true that after correcting for differences in denominations, differences in real wages are negligible. One way to think of this is that the greatest part of the variation in hourly wage comes from the relatively unimportant fact that currency is denominated differently in dif-

Country	Burger Price	Hourly Wage	Real Wage
Denmark	24.75	211.13	8.53
Australia	3.00	19.86	6.62
New Zealand	3.60	21.94	6.09
Switzerland	6.30	37.85	6.01
United States	2.54	14.32	5.64
Britain/UK	1.99	11.15	5.60
Germany	2.61	14.32	5.49
Canada	3.33	16.78	5.04
Singapore	3.30	15.65	4.74
Sweden	24.00	110.90	4.62
Hong Kong	10.70	44.26	4.14
Spain	2.37	8.59	3.62
South Africa	9.70	30.86	3.18
France	2.82	8.50	3.01
Poland	5.90	11.80	2.00
Hungary	399.00	704.34	1.77
Czech Rep.	56.00	85.34	1.52
Brazil	3.60	4.58	1.27
South Korea	3000.00	3134.00	1.04
Mexico	21.90	17.61	0.80
Thailand	55.00	31.69	0.58
China	9.90	5.56	0.56
mean	166.01	207.32	3.72
std. dev.	638.49	670.63	2.29
std. dev./mean	3.85	3.23	0.62

Table 1: Hourly Wage versus Real Wage. Burger price and hourly wage are in the local currency. Burger price is the average cost of a Big Mac. The units for Real Wage are burgers per hour. Data comes from Behar’s “Who earns the most hamburgers per hour?” [2]. The correlation coefficient between burger price and hourly wage is $\rho = 0.99$.

ferent countries. The standard deviation of hourly wages in nominal terms is about 300 times as large as that in real terms. Although the standard deviation of real wages across countries is tiny compared to that of nominal exchange rates, this variation is far more important for the quality of life of workers. Thus, one would be wrong to conclude from the high correlation coefficient that the real wage is constant across countries. Quite the contrary; the standard deviation of this ratio is 62% of the mean.

2 Eigenfactor and Total Citations

Davis (2008) fell into a similar trap in his recent comparison of journal rankings by Eigenfactor score and by impact factor or total citations [5]. In that paper, Davis aimed to determine whether measures of “popularity” such as impact factor and total citation differ substantially from measures of prestige such as the journal PageRank [4] and the Eigenfactor metrics [3]¹. To do so, Davis conducted a regression analysis of Eigenfactor scores on total citations² for a set of 165 medical journals³. Davis reports that the

¹The same issue is given more extensive analysis in ref [4]. In that paper, Bollen and colleagues compare weighted PageRank with impact factor and with total citations to explore differences between popularity and prestige. Weighted PageRank and Eigenfactor are both variants of the PageRank algorithm.

²In his paper Davis also looked at the correlation coefficient between Eigenfactor and impact factor scores. This ρ value is lower ($\rho = 0.86$), but the point is not so much what this value is, but rather that the comparison makes little sense. Eigenfactor is a measure of total citation impact, and should (all else equal) scale with the size of the journal. Impact factor is a measure of citation impact per paper, and all else equal should be independent of journal size. If one wants to compare an Eigenfactor metric with the Impact Factor, one should use the Article Influence Score, which is a per-article measure like Impact Factor.

³Contrary to what is specified in that paper, Davis appears to have sampled from both the “Medicine General and Internal” and “Medicine Research and Experimental” fields,

correlation coefficient between 2006 Eigenfactor scores and total citations⁴ is $\rho = 0.9493$. From this, Davis concluded that “*the concepts of popularity (as measured by total citation counts) and prestige (as measured by a weighting mechanism) appear to provide very similar information.*” But is Davis right? Is it really the case that if you know the number of citations, you would be wasting your time by finding the Eigenfactor score? Not at all.

The reason: there are enormous differences in the size of academic journals, which swamp the patterns that Davis was seeking in his analysis. The JCR indexes journals that range in size from tiny (*Astronomy and Astrophysics Review* has published 13 articles over the previous five years) to huge (*The Journal of Biological Chemistry* has published 31,045 articles over the same period) with a coefficient of variation, c_v , equal to 1.910. Per-article citation intensity varies less, whether measured by Article Influence or by Impact Factor (AI: range 0–27.5, coefficient of variation= 1.785; IF: range 0–63.3, coefficient of variation= 1.548).

We can formalize these observations by decomposing Davis’ regression of Eigenfactor on total citations. Davis regresses

$$\text{Log}(EF_i) \text{ vs } \text{Log}(CT_i),$$

where EF_i is the Eigenfactor score for journal i and CT_i is the total citations not merely the former category. In our analysis of the same subfields of medicine, we included 168 journals (of the 171 journals in this field); we eliminated 3 journals because they had an impact factor and/or Article Influence score of zero

⁴Davis appears to have used citations (from year 2006) to all articles published in the journals he selected. A cleaner comparison, which would have resulted in a higher correlation, would have been to extract citations (from year 2006) to articles published in the past five years, since the Eigenfactor score takes into account only the past 5 years’ citations.

received by journal i . We let AI_i be the Article Influence for journal i , and $N_{i,5}$ is the total number of articles published over the last five years for journal i . Then by definition

$$\begin{aligned}\log(EF_i) &= \log(c_1 \times AI_i \times N_{i,5}) \\ &= \log c_1 + \log AI_i + \log N_{i,5},\end{aligned}$$

where c_1 is a scaling constant that normalizes the Article Influence scores so that the mean article in the JCR has an Article Influence score of 1.00. Similarly, letting IF_i be the impact factor for journal i ,

$$\begin{aligned}\log(CT_i) &\approx \log(c_2 \times IF_i \times N_{i,2}) \\ &\approx \log(c_2 c_3 \times IF_i \times N_{i,5}) \\ &= \log c_2 c_3 + \log IF_i + \log N_{i,5}\end{aligned}$$

where c_2 and c_3 are additional scaling constants. The scaling constant, c_2 , accounts for the fact that Davis compared citations for *all* years and not just citations for 2 years. The scaling constant c_3 relates the number of articles published in two years to the number of articles published in five years (and thus is approximately 5/2). As a result, Davis is effectively calculating a regression between

$$\log(\text{Article Influence}) + \mathbf{\log(\text{total articles})}$$

and

$$\log(\text{impact factor}) + \mathbf{\log(\text{total articles})}.$$

Having the “log(total articles)” term on both sides of the regression — especially given that it varies more than the other two terms — obscures the relation between the variables that one would actually wish to observe when trying to evaluate the difference between “popularity” and “prestige”.

To better understand the consequences of having the same random variable as one of the terms on each side of the regression, an example is useful. Let X be a uniform, random variable from $1 - 10,000$ and Y and Z be uniform, random variables from $1 - 10$. Now consider the correlation coefficient between

$$\log Y + \log X \quad \text{and} \quad \log Z + \log X.$$

$\log X$ appears on both sides of this comparison, just as $\log N_{i,5}$ appears on both sides of Davis’s comparison. Even with a zero correlation between Y and Z , the correlation of this relationship is approximately 0.7. This is illustrated graphically in Figure 1. The smaller the dark box — i.e., the smaller the range of Y and Z relative to the range of X — the closer the correlation will be to one.

Here we see that even in the unlikely event that per-article popularity (as measured by Impact Factor) and per-article prestige (as measured by Article Influence score) were entirely uncorrelated, Davis still would have observed a high correlation coefficient in his regression of Eigenfactor and total citations ($\sim \rho = 0.6$ for all journals), because of the presence of $\log(\text{articles})$ on both sides. What Davis discovered is not that popularity and prestige are the same thing; it is that big journals are big and small journals are small. Because of this wide variation in journal size, one would also observe a high correlation coefficient between pages and total cites, though very few would

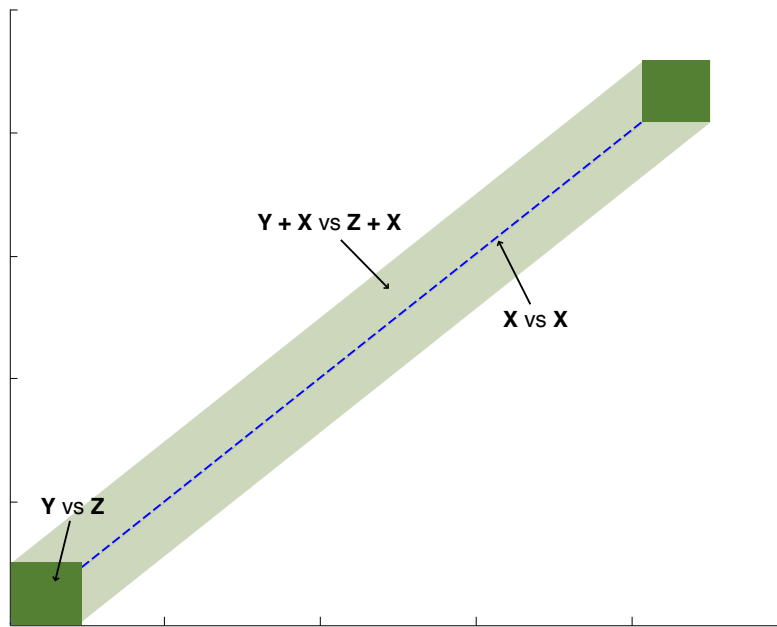


Figure 1: The correlation between $Y + X$ and $Z + X$. Even if Y and Z are independent, we see a strong correlation between the sums $Y + X$ and $Z + X$ due to the appearance of X on both sides.

argue that the former is an adequate surrogate for the latter⁵.

Another way to think about this is to focus on the *ratio* of Eigenfactor to total citations. Notice that this is a measure of “bang per cite” – that is, how much Eigenfactor boost does a citation from this journal contribute. In the hamburger example, the corresponding notion is “burgers per hour,” the real wage or purchasing power of an hour’s work. Does a high correlation between citation count and Eigenfactor score mean that the bang per cite — the ratio of Eigenfactor to citations is about constant? If it is, there really would be no point to looking at Eigenfactor scores instead of total cites. So let’s see what happens.

Figure 2 shows the ratio of Eigenfactor score to total citations for every journal in the JCR (the insert shows just the medical journals). The standard deviation of this ratio is 1.1×10^{-5} and the mean is 1.56×10^{-5} . The standard deviation, in this case, is 71% of the mean. This is even more variable than the Big Mac case! Moreover, there are nearly 1000 journals with twice the mean “bang per cite”.

The thing to notice in both the Big Mac and the journal example is that if you are interested in the ratio of A to B and if $A = ax$ and $B = bx$ for some x with a very high variance relative to that of a and of b , you will get a very high ρ value when you regress B on A . However, if what really interests you is the ratio A/B , you will note that the x ’s cancel and $A/B = ax/bx = a/b$. Thus, the variance of x has literally nothing to tell you about the variance of the ratio a/b . You learn nothing at all about whether a/b is nearly constant or highly variable from looking at the correlation of B on A .

So, if correlation coefficients are misleading, what is the alternative?

⁵We collected page and citation information for 149 Economics journals in 2006. The correlation coefficient between total pages and total citations is $\rho = 0.615$.

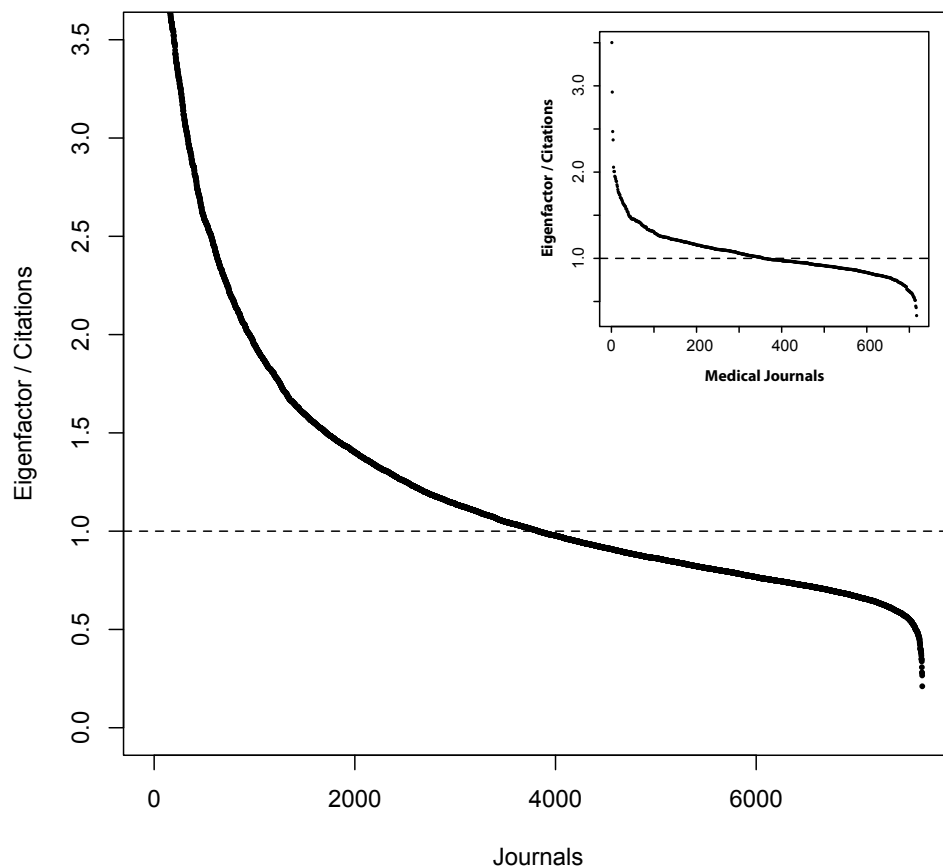


Figure 2: Ratio of Eigenfactor score to total citations. Data are normalized by the median ratio of the data set. The dashed line indicates a ratio of one. The journals are ordered from those with the highest ratio to the lowest. The inset shows only the 168 medical journals from Davis's analysis.

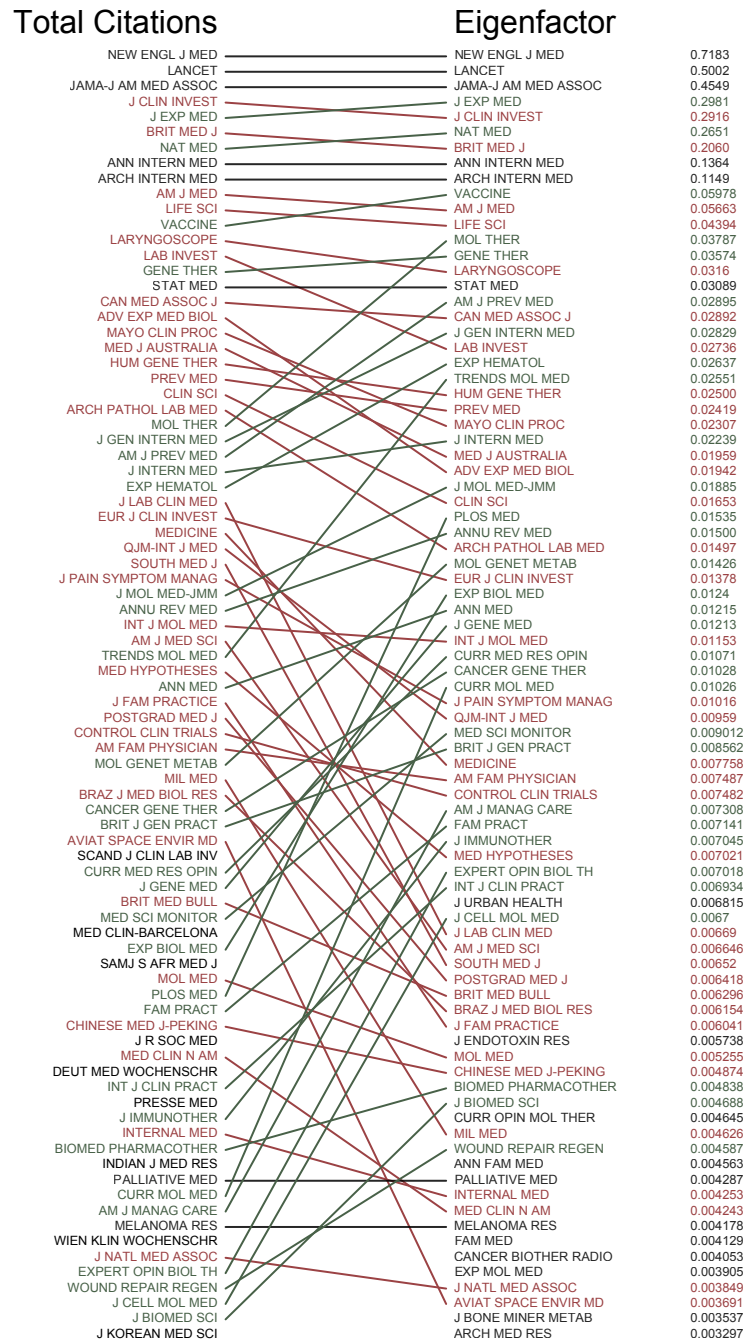
First, we argue for a deeper examination of the data. Figure 3 is an example of this strategy⁶. Listing the journals in this way, one is able to quickly see the ordinal differences that exist between this highly correlated data. This type of graphical display illustrates the interesting stories that can be lost behind a summary statistic such as the Spearman correlation.

Figure 3 illustrates the ordinal ranks of the top 50% of the medical journals used in Davis’s study. In the left column, the journals in this subfield of medicine are ranked by the total number of citations. In the right column, the journals are ordered by Eigenfactor score. The lines connecting the journals indicate whether the journal moved up (green), down (red) or stayed the same (black) relative to their ranking by total citations. The figure highlights the differences between the metrics. For example, *Aviation Space and Environmental Medicine* drops 30 places while *PLoS Medicine* raises 31 places. Davis claims in his paper that the ordering of journals does not change drastically. Figure 3 suggests otherwise.

3 Article Influence and Impact Factor

If Davis’s 2008 analysis is problematic for the various reasons described above, what is the proper way to address the relation between popularity and prestige? It would certainly help to take the appropriate regression –

⁶**Figure 3 caption:** Journal ranking comparisons by total citations and Eigenfactor score. The journals listed are the top 50% from the field of Medicine that Davis analyzed. Journals in the left column are ranked by total citations for all years. Journals in the right column are ranked by Eigenfactor score. The lines connecting the journals indicate whether the journal moved up (green), down (red) or stayed the same (black) relative to their ranking by total citations. Journal names in black can also be journals that do not exist in both columns.



between popularity *per article* and prestige *per article*. That is, we need to look at the comparison

$$\text{Log}(\text{Article Influence}) \text{ vs. } \text{Log}(\text{Impact Factor}).$$

Since its inception in January 2007, Eigenfactor.org has provided exactly this information at <http://www.eigenfactor.org/correlation/>, for the entire JCR dataset and also for each individual field of scholarship as defined by the JCR⁷. Figure 4 is a histogram of the correlation coefficients between Impact Factor and Article Influence scores for all 231 categories in the 2006 JCR. The mean for all fields was 0.853 with a standard deviation of 0.099. The field with the lowest correlation coefficient is Communication ($\rho = 0.478$). Marine Engineering has the highest correlation ($\rho = 0.986$). The sample of medical journals that Davis selected, with $\rho = 0.954$, ranks in the 90th percentile when compared to all 231 fields.

Correlation coefficients within fields typically exceed the correlation coefficient for all journals together. For all 7,611 journals considered together, $\rho = 0.818$. This value is lower than the mean of individual-field correlation coefficients, which is 0.853.

Figure 5 compares the ordinal ranking by Impact Factor and Article Influence for 84 journals (the top-ranked half) from Davis’s study . (Changes in ranking are even more dramatic when we look at the lower-ranked 84 journals)⁸. The correlation coefficient between Impact Factor and Article Influence for these 84 journals is $\rho = 0.955$. Despite this high correlation, the

⁷Falagas et. al (2008) presented a similar comparison of Impact Factor and the SJR indicator (a per-article measure of prestige) [6].

⁸**Figure 5 caption:** Comparing Impact Factor and Article Influence. The journals shown are from the same field that Davis analyzed (because of limited space, only the top 84 journals are shown). For these 84 journals, the correlation coefficient between IF and

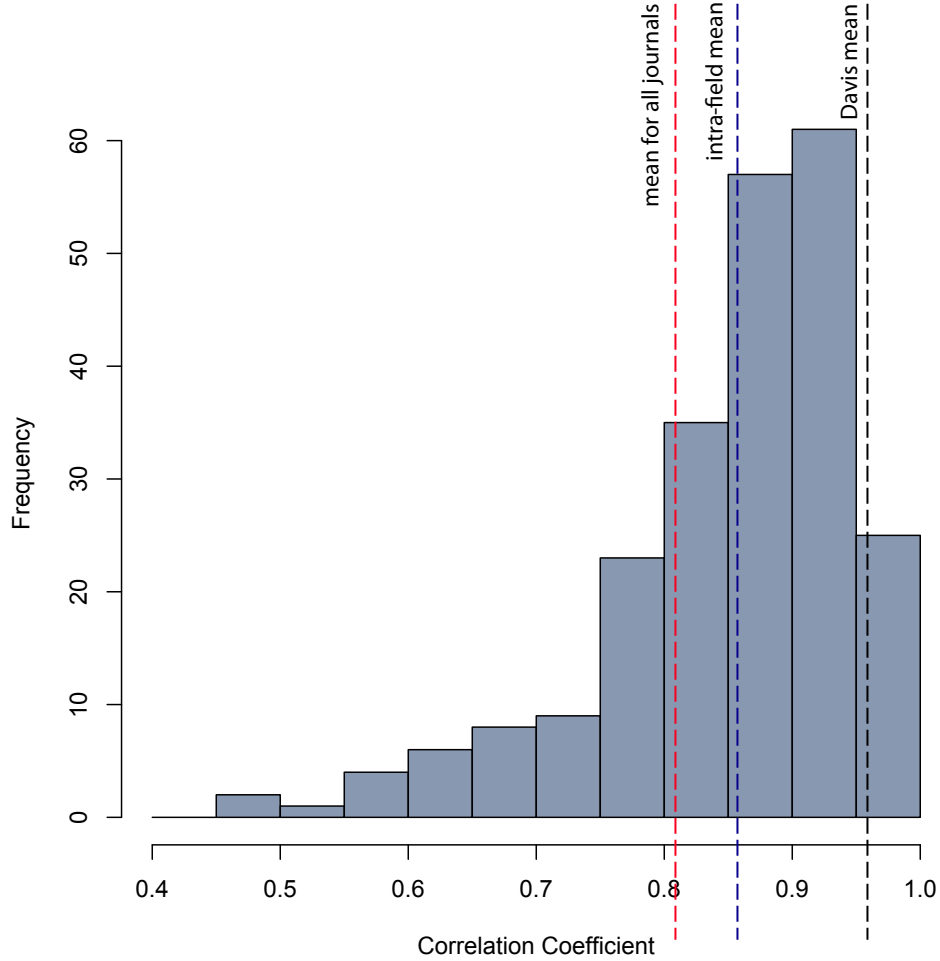


Figure 4: Histogram of correlation coefficients between Impact Factor and Article Influence scores. This includes all 231 categories in the 2006 Science and Social Science JCR. The mean of all fields is 0.853 (intra-field mean) and the standard deviation is 0.099. The correlation for all journals considered together is 0.818. The correlation for the field that Davis looked at (Medicine) is 0.954. The correlation coefficients for all fields can be found at <http://www.eigenfactor.org/correlation/>.

figure highlights the fact that the two metrics yield substantially different ordinal rankings.

Figure 5 reveals that the top few journals change in rank less than those further down the hierarchy. For example, going from Impact Factor to Article Influence, the journals in the top ten change in rank by only 1 or 2 positions. By contrast, there are many larger changes further on in the rankings⁹. For example, as we go from Impact Factor to Article Influence, the *Journal of General Internal Medicine* rises 18 spots to number 19 while *Pain Medicine* drops 35 spots to end up at number 80. These are just two of the many major shifts (in a field with a correlation of 0.955!). These changes in relative ranking would certainly not go unnoticed by editors or publishers.

Furthermore, while ordinal changes are interesting, cardinal changes are often more important. Figure 6 shows the top ten journals from Figure 3 — those with the least ordinal change from one metric to another — now in their cardinal positions. Even those journals that do not change ordinal rank from one metric to another may be valued very differently under the two different metrics. For example, *Nature Medicine* is the #2 journal regardless

AI is $\rho = 0.955$. The relative rankings by Impact Factor and Article Influence are listed in the left and right column, respectively. The third column lists the Article Influence scores. The journal names in green indicate those that fare better when ranked by Article Influence; the journal names in red fare better when ranked by Impact Factor. The names in black are journals that exhibit no change or exist outside the range of the journals shown.

⁹Bollen (2006) [3] observed a similar pattern in a series of scatterplots contrasting PageRank and Impact Factor values for all journals. In these scatterplots the rankings of top-tier journals differ relatively little whereas more variation is found in the middle and bottom portions of the hierarchy.

Impact Factor

Article Influence

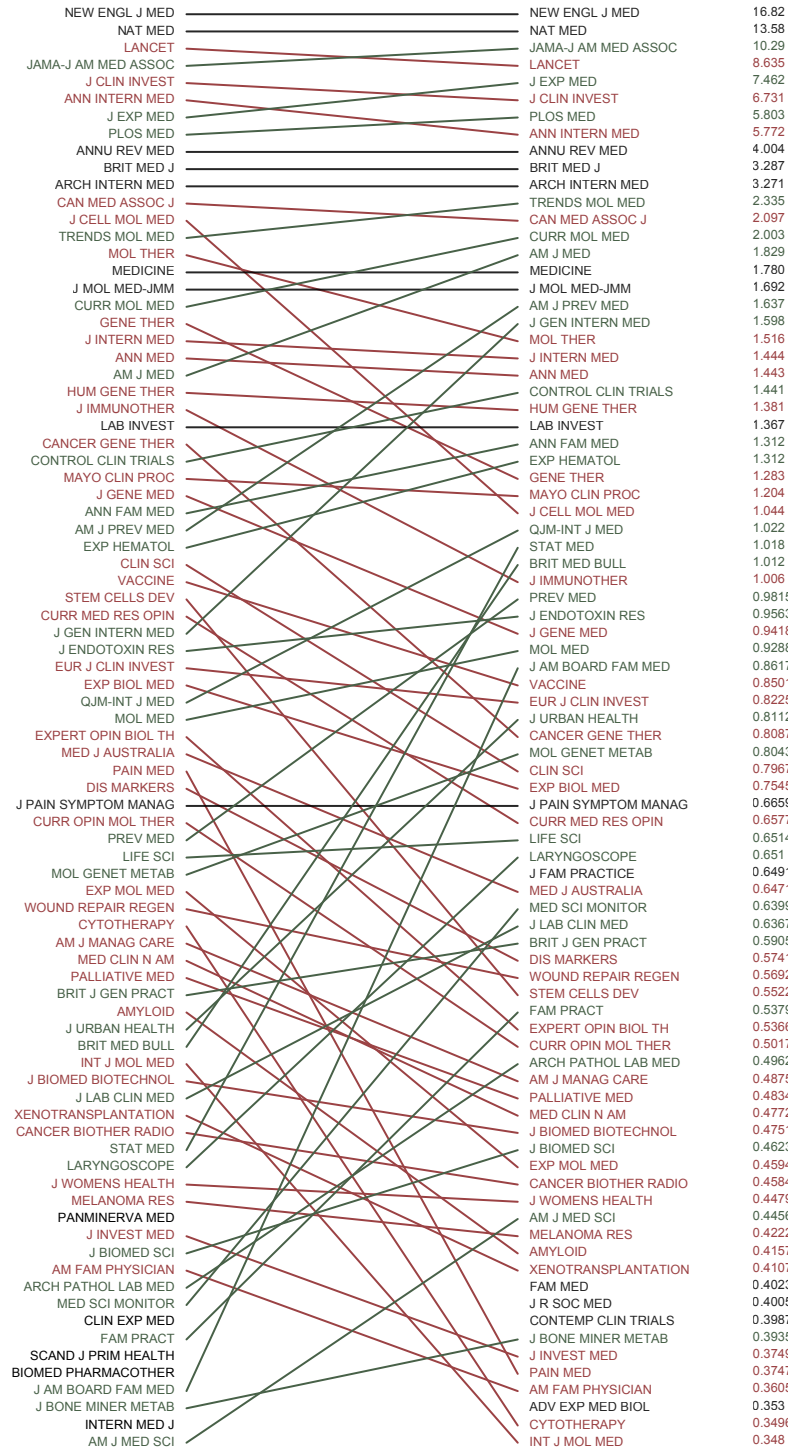


Figure 5: See footnote in text for caption.

of whether one uses Impact Factor or Article Influence. But under Impact Factor, it has barely half the prestige of the first-place *New England Journal of Medicine*, whereas by Article Influence it makes up a good deal of that ground.

Cardinal changes become especially important when comparing the value of competing journals. For example, the *Journal of Economic Theory* (JET) is ranked 7th in Eigenfactor among economics journals and the *Rand Journal of Economics* (RAND) is ranked 22nd. But a librarian making acquisitions decisions needs a measure of "how much more valuable is JET than Rand?" It turns out that JET was cited 471 times and RAND is cited 314 times, so does that mean that JET is worth 1.5 times as much as RAND? Not necessarily. JET's Eigenfactor score is about twice that of RAND. Thus, the librarian can conclude that RAND is worth about half as much as JET.

But what happens when price is included in the comparison? RAND may be worth half as much as JET, but RAND costs only a 1/10 of JET. RAND is more cost effective, both under price per citation and price per Eigenfactor. Whether a librarian chooses to use raw citations or Eigenfactor does not matter in this cost comparison. However, there are examples where the cost effectiveness is different under Eigenfactor. The journal, *American Family Physician*, received 936 citations, ranks 25th in Eigenfactor among medical journals, and costs \$204 (a better deal under citations). The *British Journal of General Practice* received 781 citations, ranked 23rd in Eigenfactor among medical journals, and costs \$220 (a better deal under Eigenfactor). If a medical librarian was comparing the cost effectiveness between these two competing journals, he or she would come to different conclusions using raw citations versus Eigenfactor scores.

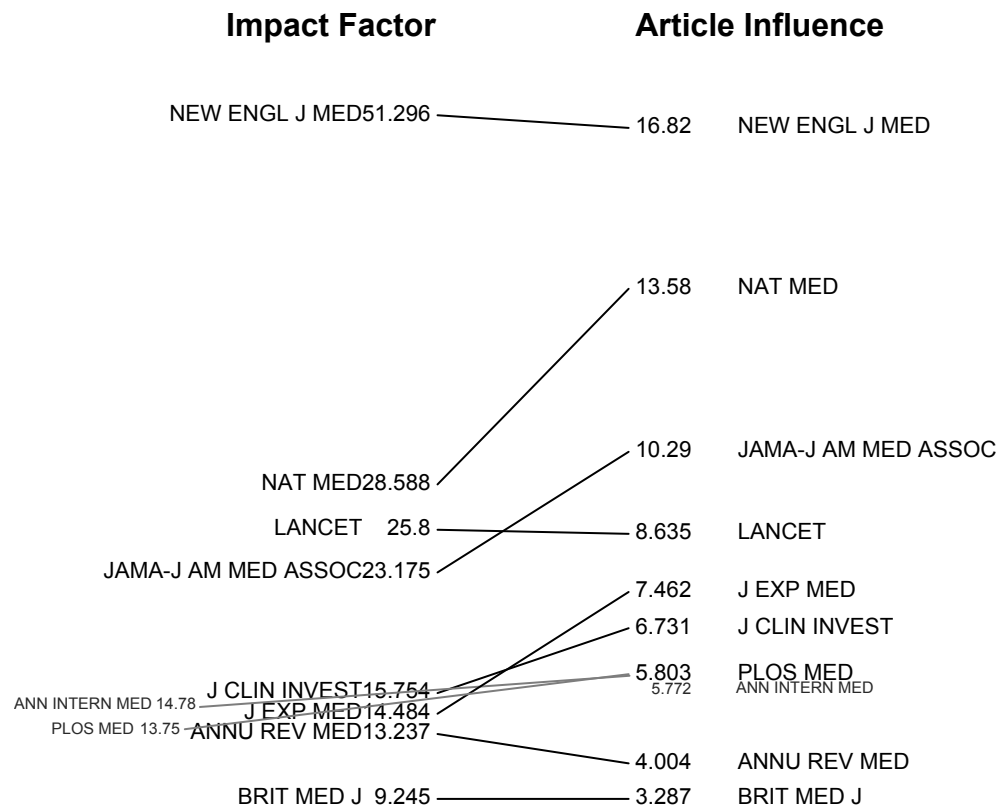


Figure 6: Cardinal differences between impact factor and Article Influence score. The top ten journals by impact factor are shown in the left column. The scores are scaled vertically, reflecting their cardinal positions. The smallest impact factor score is on the bottom, and the highest impact factor score is on the top. The right column shows the same journals scaled by Article Influence.

4 Conclusion

Correlation coefficients can be useful statistical tools. They can help us identify some kinds of statistically significant relationships between pairs of variables, and they can tell us about the sign (positive or negative) of these relationships. One must use considerably greater caution, however, when drawing conclusions from the magnitude of correlation coefficients — all the more so in the presence of spurious correlates and in the absence of any hypothesis-testing framework. In particular, we have illustrated that just because two metrics have a high correlation — 0.8 or 0.9 or even higher — we cannot safely conclude that they convey the same information, or that one has little additional information to tell us beyond what we learn from the other.

Comparative studies of alternative measures can be very useful in choosing an appropriate bibliometric toolkit. We close with a few suggestions for how one might better conduct these sorts of analyses. First, be wary of what correlation coefficients say about the relationship of two metrics [7, 1]. High correlation does *not* necessarily mean that two variables provide the same information any more than a low correlation means that two variables are unrelated. Purchasing power varies wildly despite the high correlation between wage and hamburger price in our Big Mac example. At the other end of the spectrum, in the chaotic region of the logistic map, successive iterates have an immediate algebraic relationship yet a correlation of zero.

Second, appropriate data visualization can bring out facets of the data that are obscured by summary statistics. Different forms of data graphics can be better suited for certain tasks; for example the comparison plots such as those in Figure 5 better highlight the differences between bibliometric

measures than do standard scatter plots.

Finally, simple observations can be at least as powerful as rote statistical calculations in understanding the nature of our data. For example, the median of the burgers/hour in the top third of the countries is about five times the median of the burgers/hour in the bottom third. This says a great deal about the differences in purchasing power across countries. The median “bang per cite” in the top third of journals is almost 2.4 times of the median in the bottom third. This says a great deal about the difference in how journals are valued under the Eigenfactor metrics, and helps us understand why the Eigenfactor metrics offer a substantially different view of journal prestige than that which we get from straight citation counts.

5 Acknowledgements

We would like to thank Ben Althouse for assistance with figures 3, 5, and 6, Cosma Shalizi for helpful discussions, and Johan Bollen for numerous comments on the manuscript. This research was supported in part by NSF grant SBE-0915005 to CTB.

References

- [1] FJ Anscombe. Graphs in statistical analysis. *American Statistician*, pages 17–21, 1973.
- [2] A. Behar. Who earns the most hamburgers per hour? Accessed on the web 8/8/30.

- [3] C.T. Bergstrom. Eigenfactor: Measuring the value and prestige of scholarly journals. *College and Research Libraries News*, 68(5):314–316, May 2007.
- [4] J. Bollen, M.A. Rodriguez, and H. Van de Sompel. Journal status. *Scientometrics*, 69(3):669–687, 2006.
- [5] P.M. Davis. Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts? *Journal of the American Society for Information Science and Technology*, 59(13):2186–2188, 2008.
- [6] M.E. Falagas, V.D. Kouranos, R. Arencibia-Jorge, and D.E. Karageorgopoulos. Comparison of scimago journal rank indicator with journal impact factor. *The FASEB Journal*, 22(8):2623–2628, 2008.
- [7] J.W. Tukey. Unsolved problems of experimental statistics. *Journal of the American Statistical Association*, 49(268):706 – 731, 1954.