SOME PITFALLS WITH CITATION STATISTICS

Ignacio E. Grossmann
Carnegie Mellon University • Pittsburgh, PA 15213-3890

The motivation for writing this article is to report on an experience that we in the Department of Chemical Engineering at Carnegie Mellon University had with citation statistics. We believe it is worth sharing this experience, particularly in light of the recent article by Angus, *et al.*,[1] who proposed alternative ways of measuring quality for ranking chemical engineering departments. The main idea in that article was to eliminate surveys and rely exclusively on quantitative measures, with citations being one of the major metrics. As we will describe in this article, great care has to be exercised in gathering and interpreting these data, as otherwise it is easy to obtain misleading conclusions (see Centra[2] for a general discussion on problems with citation analysis).

THE CARNEGIE MELLON CASE STUDY

In 1992, *Science Watch* [**3**(2), pp. 1-8, April 1992] published an article titled "Chemistry that Counts: The Frontrunners in Four Fields." In that article, the table on page 8 listed the following as the top six departments in citations per paper during the period of 1984-1990:

#	University	Papers 1984-90	Citations 1984-90	Citations per paper
1	Carnegie Mellon University	98	670	6.84
2	Twente University of Technology	79	490	6.20
3	University of Wisconsin, Madison	106	629	5.93
4	University of Minnesota, Minneapolis	125	697	5.58
5	University of Texas, Austin	132	732	5.55
6	Massachusetts Institute of Technology	205	1134	5.53

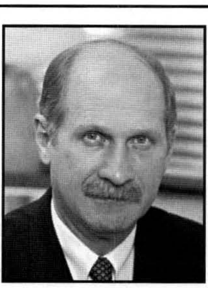
The source used in that study was 58 dedicated journals of chemical engineering (subsection of ISI Current Contents/ Engineering Technology and Applied Science).

On the other hand, according to the 1995 NRC Report, Appendix Table P (p. 500), the ranking in terms of citations per faculty for the five top U.S. departments and Carnegie Mellon for the period of 1988-92 was:

#	University	Citations 1988-92	Citations per Faculty
1	University of Minnesota, Minneapolis	3751	117.2
2	Stanford University	1039	103.9
3	University of Texas, Austin	2874	95.8
4	University of California, Berkeley	1697	89.3
5	Massachusetts Institute of Technology	2438	78.6
...			
41	Carnegie Mellon University	359	21.1

The source used was also the ISI Database and covered a considerably larger, but unspecified, number of journals.

The studies covered different periods, 1984-1990 vs. 1988-1992, as well as a different domain of journals. Nevertheless, it was clear that the number of citations reported in the NRC Report for Carnegie Mellon seemed to be much lower. In particular, the number of citations from the NRC study (359) was one-half of the *Science Watch* study (670), even though the NRC Report presumably covered a larger number of journals.



***Ignacio E. Grossmann** is the Rudolph R. and Florence Dean Professor of Chemical Engineering and Department Head at Carnegie Mellon University. He obtained his BS degree at the Universidad Iberoamericana in 1974 and his MS and PhD at Imperial College in 1975 and 1977, respectively, all in chemical engineering. His research interests are in the areas of process synthesis, energy integration, process flexibility, planning and scheduling of batch and continuous processes, and mixed-integer and logic-based optimization.*

. . . great care has to be exercised in gathering and interpreting these data, as otherwise it is easy to obtain misleading conclusions.

The above discrepancy prompted us to conduct an independent study in the summer of 1996. We received two databases from ISI containing the names and papers from our faculty in the period 1981-1995. In the first database, no biological science journals were included; in the second, they were included. The numbers that we found were as follows for the period 1988-1992 for Carnegie Mellon—we included only active faculty (17 faculty, as in NRC Report) during that period (retired or deceased faculty were excluded):

a) Without biological sciences

Total citations	1241
Citations per faculty	73

b) With biological sciences

Total citations	2747
Citations per faculty	162

The reason for the large increase in citations with biological sciences was that several papers were published by Rakesh Jain in *Cancer Research.* For instance, one of his papers[3] had a total of 265 citations for the 1981-1995 period.

So, what can we conclude from the above numbers?

Even if we were to exclude the biological science journals and remove 40% as an estimate for taking the data in 1996 rather than in 1993 (see point #3 below), the statistics are

Total citations	745
Citations per faculty	44

Therefore, compared to the NRC numbers there was at least a difference factor of two in the number of citations. In fact, if we consider the worst case (only 745 citations), Carnegie Mellon's rank would have been 14, with 44 citations per faculty. In the more realistic case of 1241 citations, our rank would have been 6, with 73 citations per faculty. In both cases, there is clearly a rather large discrepancy with the original rank of 41 for citations per faculty.

It should also be pointed out that the number of publications per faculty reported for Carnegie Mellon in Appendix K (p. 286) is significantly lower than it should be. That table reports 8.2 publications per faculty compared to 14.9 from the ISI database (*i.e.,* 254 publications and 17 faculty). Therefore, based on the count of number of publications, only about one-half were considered in the NRC Report.

We contacted both the NRC and the ISI for clarification. Based on their input, as well as on our experience in working directly with the ISI database, we summarize below the possible pitfalls that we identified with the citation statistics.

WHAT ARE THE POTENTIAL PITFALLS?

1. Misspelling of Names of Authors

This is a rather simple, but very critical, issue that we found when requesting information from ISI. The two extreme cases are 1) common names and 2) names that are easily misspelled. For example, in our department the data we received from John L. Anderson, our current Dean of Engineering, contained a very large number of papers in other areas. We had to manually separate the entries that corresponded to our John Anderson because the database did not allow simultaneous specification of both name and affiliation.

At the other extreme, the name of the author of this manuscript, Ignacio E. Grossmann, was initially misspelled with one "n" (Grossman), and a similar difficulty occurred with Andrew Gellman (Gelman). As a consequence, we received only a handful of citations in the initial request; the ones with the misspelled last names. Missing middle initials was another problem.

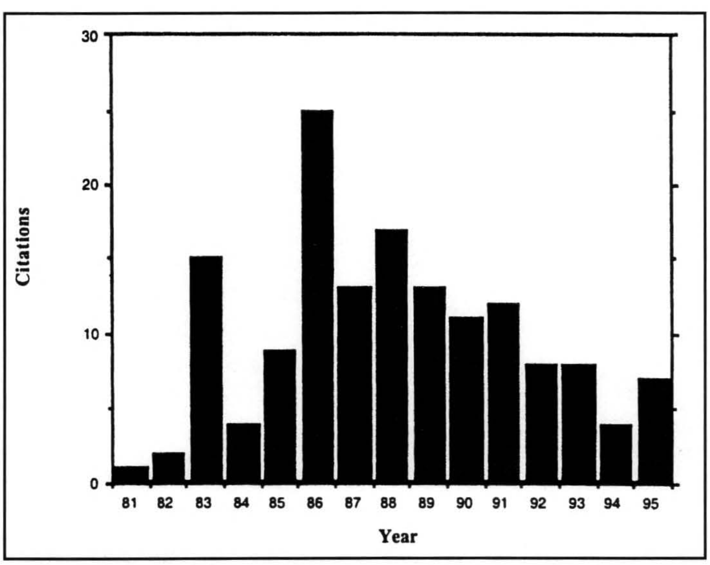


Figure 1. *Plot of citations versus time for* IEEE, ***69,*** *p. 1232 (1981)*

2. Domain of Journals for Search

As the study in the 1992 *Science Watch* indicated, a large number of journals was excluded (compared to the NRC Report) since the study was confined to "chemical engineering" journals. The NRC, however, was also not immune to problems. We were told by its staff that only certain disciplines were associated with each journal. This clearly means that departments with faculty publishing in the nontraditional disciplines were most probably penalized.

3. *Timing of Measurement of Citations Relative to Publication Times*

In our study at Carnegie Mellon, we found the following interesting observation: Most papers that have a significant number of citations (say, greater than 50), achieve their maximum number of citations between 4 to 6 years from publication. An example of this is a paper by Arthur Westerberg[4] that had a total of 149 citations. As shown in Figure 1, the maximum number of citations in this paper was in 1986, five years after its publication. This trend held in many of our papers. The implication is clear: Statistics for papers that have been out for only 1 to 3 years will probably miss more than 75% of the citations that such a paper may receive.

4. *Interpreting the Number of Citations*

One of the most important issues when reporting the number of citations over a given time period is determining what this exactly means. Intuitively, it may appear that it is the total number of citations that were made to a given author for that time period and for papers published *prior to* and during that period. But if one uses an ISI database over a given time period, the result obtained is only the number of citations of papers published in that particular time period.

To give a specific example, consider the 1988-92 time period that was used by NRC, measured in 1993. The total number of citations would intuitively be the citations of papers published before 1988-92 and during 1988-92. But what one obtains from the ISI database is only the number of citations of those papers published *during* the period 1988-92. Therefore, according to point #3 above, for a 1988 paper we pick up five years of the life of a paper, while for a 1992 paper we pick up only one year of its life. Aside from the fact that this will be an inaccurate count that will greatly underestimate the number of citations, it will be biased toward papers that are cited earlier in their lifetime (*i.e.*, papers of immediate impact).

5. *Variations of Citations by Areas*

This is a well-known fact, but it deserves discussion. Let us consider the two papers

1 Jain, R., "Determinants of Tumor Blood-Flow: A Review," *Cancer Research* (1988)
2 Fortescue, Kershenbaum, Ydstie, "Implementation of Self-Tuning Regulators with Variable Forgetting Factors," *Automatica* (1981)

Up to 1995, paper #1 had 265 citations, while paper #2 had 195 citations. Based on point #3, we might say paper #1 may still have some way to go to increase its number of citations. Furthermore, it already has more citations than paper #2. Should we then conclude paper #1 is more successful than paper #2?

Consider the following fact: In *Cancer Research* the expected number of citations of any given paper is 163.6; in *Automatica* it is only 15.9. (According to the ISI, the expected number of citations is the number of citations from papers of that journal, divided by the number of papers in that year.) If we divide the number of citations by the expected number of citations in the journal, one might argue that paper #2 is ten times more successful!

Finally, a related issue in citation statistics is the "impact score" of each journal, which often greatly varies by research area and largely has to do with the size of its audience. The impact score is calculated by dividing the number of citations in the past two years by the number of articles published during the same period. Statistics reported (URL: http://fellini.sissa.it/~furio/journal.html) in 1996 for some journals where chemical engineers publish are

1.359	*AIChE Journal*
1.056	*Industrial & Engineering Chemistry Research*
0.902	*Chemical Engineering Science*
0.532	*Canadian Journal of Chemical Engineering*
0.488	*Chemical Engineering Research & Design*
0.385	*Chemical Engineering Communications*
25.466	*Nature*
22.067	*Science*
22.524	*Pharmacological Reviews*
12.48	*Journal of Cell Biology*
7.507	*Journal of Clinical Oncology*
1.228	*Physics Letters A*
3.056	*Physics Letters B*
6.626	*Physical Review Letters*
3.635	*Journal of Chemical Physics*
2.492	*Journal of Catalysis*
2.745	*Surface Science*
1.864	*Journal of Fluid Mechanics*
3.016	*Macromolecules*
3.232	*Langmuir*
1.401	*Colloids and Surfaces*
1.62	*Journal of Colloid and Interface Science*
2.603	*Environmental Science & Technology*
0.9	*Automatica*
0.641	*Computers & Chemical Engineering*
0.864	*Mathematics of Operations Research*
0.763	*Mathematical Programming*
0.729	*Operations Research*
0.356	*European Journal of Operational Research*

Based on the above, it is clear that the impact factors are quite uneven in different areas. Furthermore, it is interesting to note that the *Journal of Fluid Mechanics* and *Mathematical Programming,* two journals that are notoriously difficult for accepting papers, have rather low impact factors compared to *Nature* and *Science*, which are also very selective. These observations would suggest the need for some type of normalization; *e.g.,* 30 citations of an article in the *Journal of Fluid Mechanics* could be considered a success, while 30 in *Nature* would correspond to an average paper in that journal.

CONCLUDING REMARKS

This article has demonstrated, using the experience at Carnegie Mellon with statistics on the numbers of citations, that there are a number of important potential pitfalls in compiling that type of information due to the great complexity in gathering and interpreting the information. Our experience with *Science Watch* and the NRC Report suggest that there is a pressing need for organizations that perform department rankings to carefully and rationally define measures of citations in order to avoid errors, misinterpretations, and biases against certain research areas. While we do not attempt to offer specific remedies, it would seem that the following five general policies merit consideration:

1. *Develop an identification number for authors to avoid problems with misspellings and duplicate names.*
2. *Expand the domain of search to all areas to avoid penalizing authors who publish outside of their discipline.*
3. *Consider normalizing citations according to the impact score, to reduce discrepancies between different research areas.*
4. *Ensure that the number of citations over a given time period cover publications before and during that time period.*
5. *Consult with departments to verify the statistics before publishing them, to allow for possible corrections of mistakes.*

In order to implement these policies, close collaboration with ISI and agencies performing the rankings is required. It is also hoped that this article will temper the enthusiasm of those who apply sophisticated analysis tools to questionable data, and thereby are likely to draw incorrect conclusions.

REFERENCES

1. Angus, J.C., R.V. Edwards, and B.D. Schulz, "Ranking Graduate Programs: Alternative Measures of Quality," *Chem. Eng. Ed.,* **33**(1), 72 (1999)
2. Centra, J.A., *Reflective Faculty Evaluation,* Jossey-Bass, California (1993)
3. Jain, R., "Determinations of Tumor Blood-Flow: A Review," *Cancer Research* (1988)
4. Westerberg, Arthur, *IEEE,* **69**, 1232 (1981) ❐

ChE letter to the editor

Dear Editor

Alves, et al.,[1] are to be congratulated on their elegant experiment on the drainage of liquids from vertical tubes of diameter 19 and 32 mm. But their paper should have stipulated that the diameter of tube used should be no less than about 15 mm. The reason is that the simple equation for the bubble velocity

$$U = 0.345(gD)^{0.5} \qquad (1)$$

does not apply for smaller tubes because of surface tension effects. The relevant dimensionless group for such effects is the Eötvos number, which is given by

$$Eo = gD^2 \Delta\rho / \sigma \qquad (2)$$

where D = tube internal diameter, $\Delta\rho$ = density difference between heavy and light phases, and σ = surface or interfacial tension.

At values of Eo less than about 50 in the case of low-viscosity liquids, the velocity U is below the value predicted from Eq. (1). Moreover, if Eo is below a critical value of 3.37, corresponding to a tube diameter of about 5 mm for air/water, the bubble will not rise at all.[2] Some years ago we measured drainage rates from vertical tubes of between 6 and 10 mm diameter, with the bubble nose (meniscus) being controlled at a constant position.[3] This technique enabled the surface tension in gas-liquid and liquid-liquid systems to be measured continuously.

M.H.I Baird, *Professor*
N.V. Rama Rao, *Research Associate*
McMaster University

REFERENCES

1. Alves, M.A., M.F.R. Pinto, and J.R.F.G. de Carvalho, "Two Simple Experiments for the Fluid-Mechanics and Heat-Transfer Laboratory Class," *Chem. Eng. Ed.,* **33**, 226 (1999)
2. Bretherton, F.P., "The Motion of Long Bubbles in Tubes," *J. Fluid Mech.,* **10**, 166 (1961)
3. Rao, N.V.R., and M.H.I. Baird, "Continuous Measurement of Surface and Interfacial Tension by Stationary Slug Method," *Can. J. Chem. Eng.,* **61**, 581 (1983) ❐